UNIVERSAL
LIBRARY

OU_164832

UNIVERSAL
LIBRARY

OUP—786—13-6-75—10,000.

OSMANIA UNIVERSITY LIBRARY

Call No. 510.001311 W64M Accession No. 28309

Author Wilks, S.S.

Title Mathematical Statistics.

This book should be returned on or before the date last marked below.

MATHEMATICAL STATISTICS

By

S. S. WILKS

UNIV

MATHEMAT[illegible] [illegible]ISTICS

By S. S. Wilks

ERRATA

Page	Line	In Place of	Read
6	15	(2)	(2')
9	4	$x' \longrightarrow -\infty$	$x_1' \longrightarrow -\infty$
16	2	E	E_1
18	6	add	and
22	1b*	$f(x_1,x_2,\ldots,x_k)$	$F(x_1,x_2,\ldots,x_k)$
31	3	δ_x	$\delta\sigma_x$
42	9	$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}$	$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}$
45	10	§2.93.	§2.92.
50	1	$\delta\frac{pq}{n}$	$\delta\sqrt{\frac{pq}{n}}$
50	2	$\frac{n}{pq}$	$\sqrt{\frac{n}{pq}}$
50	3	$\frac{p^2q^2}{n^2\lambda^2}$	$\frac{pq}{n\lambda^2}$
50	13	$\frac{1}{4n^2\lambda^2}$	$\frac{1}{4n\lambda^2}$
50	14	$\frac{1}{400}$	$\frac{1}{4}$
50	15	$\frac{1}{400}$	$\frac{1}{4}$
55	12	$(k+x)(k+x-1)\ldots\ldots(k+1)$	$(k+x-1)(k+x-2)\ldots k$
55	13	$(1+\frac{x}{k})(1+\frac{x-1}{k})\ldots\ldots(1+\frac{1}{k})$	$(1+\frac{x-1}{k})(1+\frac{x-2}{k})\ldots\ldots 1$
55	19	$(1+d)^{-\frac{d}{h}}$	$(1+d)^{-\frac{h}{d}}$
56	1b	$\sqrt{h}$	h
56	2b	$\int_0^{\rho}$	$\int_0^{2\pi}$
57	2 3 5 7	$\sqrt{\frac{h}{\pi}}$	$\frac{h}{\sqrt{\pi}}$
69	4	$\sqrt{\pi}$	$\sqrt{2\pi}$
73	top	Distribution	Distributions
79	3b	vandom	random
81	3	$\delta\sigma/n$	$\delta\sigma/\sqrt{n}$
81	4	$\delta\sigma/n$	$\delta\sigma/\sqrt{n}$
81	6	$\sigma^2/n^2\epsilon^2$	$\sigma^2/n\epsilon^2$
81	7b	$n^{-\frac{1}{2}}$	$n^{-\frac{1}{2}}\sigma$
89	10b	$(\Sigma a_1)(\Sigma b_1)$	$(\Sigma a_1^2)(\Sigma b_1^2)$
90	13	$n-r_k-1$	$n-r_k$
90	16 (2)	$n-r_k-1$	$n-r_k$
90	21 (2)	$n-r_k-1$	$n-r_k$
91	1b	f^{-2}	$\bar{f}^2$
92	5	$f(x_0)$	$f(\tilde{x}_0)$
93	3b	130	473
94	15 (2)	$n-r_k-1$	$n-r_k$
95	6	$h+1_\delta$	$h+1\delta$
98	7	distribution	distributions
106	1	$A_{\alpha\beta}$	$a_{\alpha\beta}$
107	3b	n_1	$\sum_1^k n_1$
116	4b	$(x_{1\alpha}-a_1)(x_{j\alpha}-a_j)$	$A_{1j}(x_{1\alpha}-a_1)(x_{j\alpha}-a_j$
119	1	$(\frac{n-1}{2}+1)$	$\Gamma(\frac{n-1}{2}+1)$
119	5	$\Gamma(\frac{n-1}{2}+1)$	$\Gamma(\frac{n-1}{2}+1)$
120	4	$\sqrt{a_{11}a_{12}}\ A_{12}$	$\sqrt{a_{11}a_{22}}\cdot A_{12}$

* As counted from the bottom

ERRATA

Page	Line	In Place of	Read
120	3b	$n(\bar{x}-a)$	$\sqrt{n}(\bar{x}-a)$
121	1	$(i=1,2)$	$(i=1,2,\ldots,k)$
121	4	characteristic	moment generating
128	1b	(c)	(e)
136	2	$-\frac{1}{2}(\bar{x}-a)^2$	$-\frac{n}{2}(\bar{x}-a)^2$
137	6b	$(-\frac{1}{n}\frac{\partial^3 \log P}{\partial\theta^3})_{\hat{\theta}}$	$(-\frac{1}{n}\frac{\partial^3 \log P}{\partial\theta^3})_{\tilde{\theta}}$
148	10b	significant	significance
152	1b	w	ω
157	2b	dy	dy_α
158	8	$\bar{y} + \hat{b}\bar{x}$	$\bar{y} - \hat{b}\bar{x}$
162	1b	$\phi(\theta_1\theta_2)$	$\phi(\theta_1,\theta_2)$
165	2 (3)	$(1 - \frac{N}{M})$	$(1 + \frac{N}{M})$
165	3	(M-N)	(M+N)
166	19	the likelihood	is the likelihood
167	2	y	y_α
167	7	$\dot{a}_u$	$\hat{a}_u$
168	1b	(i)	(j)
174	18	$_uC_{uq}$	λ_uC_{uq}
176	2b	$\frac{1}{n}\sum_{\alpha=1}^{n}$	$\frac{1}{n_p}\sum_{\alpha}^{n_p}$
181	1b	$Y_{.j.}$ and $Y_{..k}$	$\bar{Y}_{.j.}$ and $\bar{Y}_{..k}$
183	6	$Y_{ij.}$, $Y_{i.k}$,	$\bar{Y}_{ij.}$, $\bar{Y}_{i.k}$,
183	7	$Y_{.jk}$, $Y_{i..}$, $Y_{.j.}$, $Y_{..k}$,	$\bar{Y}_{.jk}$, $\bar{Y}_{i..}$, $\bar{Y}_{.j.}$, $\bar{Y}_{..k}$,
185	1b	assumed	not assumed
186	1	different from zero	zero
186	8b	$S_{o..}$	$S^{o}_{...}$
188	4	minimizing	maximizing
192	1b	$i = 1, 2, \ldots, s$	$i = 1, 2, \ldots, r$
193	8	the R_i are	the C_j are
194	5	$r + 1$	$r + 2$
197	14	$R_{1,\omega}$	$\hat{R}_{1,\omega}$
201	5b	$P(r_{ij})$	$p(r_{ij})$
204	9	from 1 to	from
204	9	$\frac{n-1}{s-1}$	$\frac{n}{s-1}$ to n
207	1	(3n+3)	(n+1)
207	16	$Pr(u \lesssim u^o) \gtrsim \epsilon$	$Pr(u \lesssim u^o) \lesssim \epsilon$
210	1	N	n
210	13	$(-1)^g$	$(-1)^{n-h-g}$
210	15	$(-1)^g$	$(-1)^{n-h-g}$
215	11	$n^5_{i.}!$	$n_{i.}!$
220	1	in terms	in terms of
221	10b	$P_{m,n;pN,n}$	$P_{m,n;pN,N}$
223	10	maximizing p,	maximizing $\tilde{p}$,
225	12	of N_1	of N
226	5	(§5.6)	(§5.5)
226	7	(§5.12)	(§5.5)
227	12	dx_1	$dx_{1\alpha}$
229	5b**	$x^2_{1\alpha}$	x^2_α
238	5	a_{1j}	$\lvert a_{1j}\rvert$
245	13	y_k	y_s
254	1	polynimial	polynomial
256	11b	indeterminant	indeterminate
257	19	$\lVert A^{ip}\rvert$	$\lVert A^{ip}\rVert$
258	13	(i)	i
258	3b	j-th column	i-th column
260	17	The canonical correlation	The correlation
271	19	Cochran, G. C.	Cochran, W. G.
273	18	Valewis	Valeurs

** As counted from the bottom of footnote.

MATHEMATICAL STATISTICS

By

S. S. WILKS

PRINCETON UNIVERSITY PRESS
Princeton, New Jersey

1947

Copyright, 1943, by
PRINCETON UNIVERSITY PRESS

PREFACE

Most of the mathematical theory of statistics in its present state has been developed during the past twenty years. Because of the variety of scientific fields in which statistical problems have arisen, the original contributions to this branch of applied mathematics are widely scattered in scientific literature. Most of the theory still exists only in original form.

During the past few years the author has conducted a two-semester course at Princeton University for advanced undergraduates and beginning graduate students in which an attempt has been made to give the students an introduction to the more recent developments in the mathematical theory of statistics. The subject matter for this course has been gleaned, for the most part, from periodical literature. Since it is impossible to cover in detail any large portion of this literature in two semesters, the course has been held primarily to the basic mathematics of the material, with just enough problems and examples for illustrative and examination purposes.

Except for Chapter XI, the contents of the present set of notes constitute the basic subject matter which this course was designed to cover. Some of the material in the author's Statistical Inference (1937) has been revised and included. In writing up the notes an attempt has been made to be as brief and concise as possible and to keep to the mathematics with a minimum of excursions into applied mathematical statistics problems.

An important topic which has been omitted is that of characteristic functions of random variables, which, when used in Fourier inversions, provide a direct and powerful method of determining certain sampling distributions and other random variable distributions. However, moment generating functions are used; they are more easily understood by students at this level and are almost as useful as characteristic functions as far as actual applications to mathematical statistics are concerned. Many specialized topics are omitted, such as intraclass, tetrachoric and other specialized correlation problems, semi-invariants, renewal theory, the Behrens-Fisher problem, special transformations of population parameters and random variables, sampling from Poisson populations, etc. It is the experience of the author that an effective way for handling many of these specialized topics is to formulate them as problems for the students. If and when the present notes are revised and issued in permanent form, such problems will be inserted at the ends of sections and chapters. In the meantime, criticisms, suggestions, and notices of errors will be gratefully received from readers.

Finally, the author wishes to express his indebtedness to Dr. Henry Scheffé, Mr. T. W. Anderson, Jr. and Mr. D. F. Votaw, Jr. for their generous assistance in preparing these notes. Most of the sections in the first seven chapters and several sections in Chapters X and XI were prepared by these men, particularly the first two. Thanks are due Mrs. W. M. Weber for her painstaking preparation of the manuscript for lithoprinting.

S. S. Wilks.

Princeton, New Jersey
April, 1943.

TABLE OF CONTENTS

CHAPTER I. INTRODUCTION 1

CHAPTER II. DISTRIBUTION FUNCTIONS

§2.1 Cumulative Distribution Functions 5
§2.11 Univariate Case 5
§2.12 Bivariate Case 8
§2.13 k-Variate Case 11
§2.2 Marginal Distributions 12
§2.3 Statistical Independence 13
§2.4 Conditional Probability 15
§2.5 The Stieltjes Integral 17
§2.51 Univariate Case 17
§2.52 Bivariate Case 20
§2.53 k-Variate Case 21
§2.6 Transformation of Variables 23
§2.61 Univariate Case 24
§2.62 Bivariate Case 24
§2.63 k-Variate Case 28
§2.7 Mean Value 29
§2.71 Univariate Case ; Tchebycheff's Inequality 30
§2.72 Bivariate Case 31
§2.73 k-Variate Case 32
§2.74 Mean and Variance of a Linear Combination of Random Variables 33
§2.75 Covariance and Correlation between two Linear Combinations of Random Variables 34
§2.76 The Moment Problem 35
§2.8 Moment Generating Functions 36
§2.81 Univariate Case 36
§2.82 Multivariate Case 39
§2.9 Regression 40
§2.91 Regression Functions 40

§2.92 Variance about Regression Functions 41
§2.93 Partial Correlation 42
§2.94 Multiple Correlation 42

CHAPTER III. SOME SPECIAL DISTRIBUTIONS

§3.1 Discrete Distributions 47
§3.11 Binomial Distribution 47
§3.12 Multinomial Distribution 50
§3.13 The Poisson Distribution 52
§3.14 The Negative Binomial Distribution 54
§3.2 The Normal Distribution 56
§3.21 The Univariate Case 56
§3.22 The Normal Bivariate Distribution 59
§3.23 The Normal Multivariate Distribution 63
§3.3 Pearson System of Distribution Functions 72
§3.4 The Gram-Charlier Series 76

CHAPTER IV. SAMPLING THEORY

§4.1 General Remarks 79
§4.2 Application of Theorems on Mean Values to Sampling Theory 80
§4.21 Distribution of Sample Mean 81
§4.22 Expected Value of Sample Variance 83
§4.3 Sampling from a Finite Population 83
§4.4 Representative Sampling 86
§4.41 Sampling when the p_i are known 87
§4.42 Sampling when the σ_i are also known 88
§4.5 Sampling Theory of Order Statistics 89
§4.51 Simultaneous Distribution of any k Order Statistics 89
§4.52 Distribution of Largest (or Smallest) Variate 91
§4.53 Distribution of Median 91
§4.54 Distribution of Sample Range 92
§4.55 Tolerance Limits 93
§4.6 Mean Values of Sample Moments when Sample Values are Grouped; Sheppard Corrections 94
§4.7 Appendix on Lagrange's Multipliers 97

CHAPTER V. SAMPLING FROM A NORMAL POPULATION

§5.1 Distribution of Sample Mean 98
§5.11 Distribution of Difference between Two Sample Means 100
§5.12 Joint Distribution of Means in Samples from a Normal Bivariate Distribution 100
§5.2 The χ^2-distribution 102
§5.21 Distribution of Sum of Squares of Normally and Independently Distributed Variables 102
§5.22 Distribution of the Exponent in a Multivariate Normal Distribution . . 103
§5.23 Reproductive Property of χ^2-Distribution 105
§5.24 Cochran's Theorem 105
§5.25 Independence of Mean and Sum of Squared Deviations from Mean in Samples from a Normal Population 108
§5.3 The "Student" t-Distribution 110
§5.4 Snedecor's F-Distribution 113
§5.5 Distribution of Second Order Sample Moments in Samples from a Bivariate Normal Distribution 116
§5.6 Independence of Second Order Moments and Means in Samples from a Normal Multivariate Distribution 120

CHAPTER VI. ON THE THEORY OF STATISTICAL ESTIMATION

§6.1 Confidence Intervals and Confidence Regions 122
§6.11 Case in which the Distribution Depends on only one Parameter 122
§6.12 Confidence Limits from Large Samples 127
§6.13 Confidence Intervals in the Case where the Distribution Depends on Several Parameters 130
§6.14 Confidence Regions 132
§6.2 Point Estimation; Maximum Likelihood Statistics 133
§6.21 Consistency 133
§6.22 Efficiency 134
§6.23 Sufficiency 135
§6.24 Maximum Likelihood Estimates 136
§6.3 Tolerance Interval Estimation 142
§6.4 The Fitting of Distribution Functions 145

CHAPTER VII. TESTS OF STATISTICAL HYPOTHESES

§7.1 Statistical Tests Related to Confidence Intervals 147
§7.2 Likelihood Ratio Tests 150

§7.3 The Neyman-Pearson Theory of Testing Hypotheses 152

CHAPTER VIII. NORMAL REGRESSION THEORY

§8.1 Case of One Fixed Variate 157

§8.2 The Case of k Fixed Variates 160

§8.3 A General Normal Regression Significance Test 166

§8.4 Remarks on the Generality of Theorem (A), §8.3 171

§8.41 Case 1 171

§8.42 Case 2 172

§8.43 Case 3 173

§8.5 The Minimum of a Sum of Squares of Deviations with Respect to Regression Coefficients which are Subject to Linear Restrictions 174

CHAPTER IX. APPLICATIONS OF NORMAL REGRESSION THEORY TO ANALYSIS OF VARIANCE PROBLEMS

§9.1 Testing for the Equality of Means of Normal Populations with the Same Variance 176

§9.2 Randomized Blocks or Two-way Layouts 177

§9.3 Three-way and Higher Order Layouts; Interaction 181

§9.4 Latin Squares 186

§9.5 Graeco-Latin Squares 190

§9.6 Analysis of Variance in Incomplete Layouts 192

§9.7 Analysis of Covariance 195

CHAPTER X. ON COMBINATORIAL STATISTICAL THEORY

§10.1 On the Theory of Runs 200

§10.11 Case of Two Kinds of Elements 200

§10.12 Case of k Kinds of Elements 205

§10.2 Application of Run Theory to Ordering Within Samples 206

§10.3 Matching Theory 208

§10.31 Case of Two Decks of Cards 208

§10.32 Case of Three or More Decks of Cards 212

§10.4 Independence in Contingency Tables 213

§10.41 The Partitional Approach 213

§10.42 Karl Pearson's Original Chi-Square Problems and its Application to Contingency Tables 217

§10.5 Sampling Inspection 220

§10.51 Single Sampling Inspection 221

§10.52 Double Sampling Inspection 224

CHAPTER XI. AN INTRODUCTION TO MULTIVARIATE STATISTICAL ANALYSIS

§11.1 The Wishart Distribution 226
§11.2 Reproductive Property of the Wishart Distribution 232
§11.3 The Independence of Means and Second Order Moments in Samples from a Normal Multivariate Population 233
§11.4 Hotelling's Generalized "Student" Test 234
§11.5 The Hypothesis of Equality of Means in Multivariate Normal Populations . 238
§11.6 The Hypothesis of Independence of Sets of Variables in a Normal Multivariate Population 242
§11.7 Linear Regression Theory in Normal Multivariate Populations 245
§11.8 Remarks on Multivariate Analysis of Variance Theory 250
§11.9 Principal Components of a Total Variance 252
§11.10 Canonical Correlation Theory 257
§11.11 The Sampling Theory of the Roots of Certain Determinantal Equations . . . 260
§11.111 Characteristic Roots of One Sample Variance-covariance Matrix . . 261
§11.112 Characteristic Roots of the Difference of Two Sample Variance-covariance Matrices 265
§11.113 Distribution of the Sample Canonical Correlations 268
LITERATURE FOR SUPPLEMENTARY READING 271
INDEX 279

CHAPTER I

INTRODUCTION

Modern statistical methodology may be conveniently divided into two broad classes. To one of these classes belongs the routine collection, tabulation, and description of large masses of data *per se*, most of the work being reduced to high speed mechanized procedures. Here elementary mathematical methods such as percentaging, averaging, graphing, etc. are used for condensing and describing the data *as it is*. To the other class belongs a methodology which has been developed for making predictions or drawing inferences, from a given set or *sample* of observations about a larger set or *population* of potential observations. In this type of methodology, we find the mathematical methods more advanced, with the theory of probability playing the fundamental role. In this course, we shall be concerned with the mathematics of this second class of methodology. It is natural that these mathematical methods should embody assumptions and operations of a purely mathematical character which correspond to properties and operations relating to the actual observations. The test of the applicability of the mathematics in this field as in any other branch of applied mathematics, consists in comparing the predictions as calculated from the mathematical model with what actually happens experimentally.*

Since probability theory is fundamental in this branch of mathematics, we should examine informally at this point some notions which at least suggest a way of setting up a probability theory. As far as the present discussion is concerned, perhaps the best approach is to examine a few simple empirical situations and see how we would proceed to idealize and to set up a theory. Suppose a die is thrown successively. If we denote by X the number of dots appearing on the upper face of the die, then X will take on one of the values 1, 2, 3, 4, 5, 6 at each throw. The variable X jumps from

---

*For an example of such a comparison, see Ch. 5 of Bortkiewicz' *Die Iterationen*, Springer, Berlin, 1917.

value to value as the die is thrown successively, thus yielding a sequence of numbers which appear to be quite haphazard or erratic in the order in which they occur. A similar situation holds in tossing a coin successively where X is the number of heads in a single toss. In this case a succession of tosses will yield a haphazard sequence of 0's and 1's. Similarly, if X is the blowing time in seconds of a fuse made under a given set of specifications, then a sequence, let us say of every N^{th} fuse from a production line will yield a sequence of numbers (values of X) which will have this characteristic of haphazardness or randomness if there is nothing in the manufacturing operations which will cause "peculiarities" in the sequence, such as excessive high or low values, long runs of high or low values, etc. We make no attempt to define randomness in observed sequences, except to describe it roughly as the erratic character of the fluctuations usually found in sequences of measurements on operations repeatedly performed under "essentially the same circumstances", as for example successively throwing dice, tossing coins, drawing chips from a bowl, etc. In operations such as taking fuses from a production line and making some measurement on each fuse (e. g. blowing time) the resulting sequence of measurements frequently has "peculiarities" of the kind mentioned above, thus lacking the characteristic of randomness. However, it has been found that frequently a state of randomness similar to that produced by rolling dice, drawing chips from a bowl, etc., can be obtained in such a process as mass production by carefully controlling the production procedure.*

Now let us see what features of these empirical sequences which arise from "randomizing" processes can be abstracted into a mathematical theory--probability theory. If we take the first n numbers in an empirical sequence of numbers X_1, X_2, X_3, . . . , X_n, ... , there will be a certain fraction of them, say $F_n(x)$, less than or equal to x, no matter what value of x is taken. For each value of x, $0 \leq F_n(x) \leq 1$. We shall refer to $F_n(x)$ as the empirical cumulative distribution function of the numbers $X_1, X_2, X_3, \ldots, X_n \ldots$. As x increases, $F_n(x)$ will either increase or remain constant. It is a matter of experience that as n becomes larger and larger $F_n(x)$ becomes more and more stable, appearing to approach some limit, say $F_\infty(x)$ for each value of x.

*Shewhart has developed a statistical method of quality control in mass production engineering which is essentially a practical empirical procedure for approximating a state of randomness (statistical control, to use Shewhart's term) for a given measurement in a sequence of articles from a production line, by successively identifying and eliminating causes of peculiarities in the sequence back in the materials and manufacturing operations.

If any subsequence of the original sequence is chosen "at random"(i.e.according to any rule which does not depend on the values of the X's) then a corresponding $F_n(x)$ can be defined for the subsequence, and again we know from experience that as n increases, $F_n(x)$ for the subsequence appears to approach the same limit for each value of x as in the original sequence.

Entirely similar experimental evidence exists for situations in which the empirical sequences are sequences of pairs, triples, or sets of k numbers, rather than sequences of single numbers. For example, a sequence of throws of pairs of dice would give rise to a sequence of pairs of numbers; the resistance, capacity, and inductance of each relay in a sequence of telephone relays from a carefully controlled production line would yield a sequence of triples of measurements. In considering a random sequence of pairs of numbers (X_{11}, X_{21}), (X_{12}, X_{22}), . . . , (X_{1n}, X_{2n}), . . . , we can let $F_n(x_1,x_2)$ be the proportion of pairs in the first n pairs in which the value of X_1 is less than or equal to x_1 and the value of X_2 is less than or equal to x_2. We need not list all of the properties of $F_n(x_1, x_2)$, for they are straightforward extensions of those of $F_n(x)$ considered above. The important point here is that as n increases, experience indicates that $F_n(x_1, x_2)$ appears to approach some limit $F_\infty(x_1, x_2)$ for each value of x_1 and of x_2.

In particular, suppose we group the numbers of an empirical random sequence $X_1, X_2, \ldots, X_n, \ldots$ (with empirical cumulative distribution function $F_n(x)$) into pairs (or samples of two numbers), so as to make a new sequence of pairs of numbers (X_1, X_2), (X_3, X_4), . . . , (X_{2n-1}, X_{2n}), ... As before, we have an empirical cumulative distribution function $F_n(x_1, x_2)$ for this sequence of pairs. It is an experimental fact that as n becomes larger and larger, $F_n(x_1, x_2)$ behaves more and more nearly like the product $F_n(x_1)\, F_n(x_2)$. A similar situation is true for sequences of samples of three or more numbers. As we shall see later, it is this product property that suggests a way to set up a mathematical theory of sampling.

The matter of $F_n(x)$ appearing to approach some function $F_\infty(x)$ as n increases is purely an empirical phenomenon, and not a mathematical one, but it suggests a way of setting up a mathematical model corresponding to any randomizing process which, upon repeated application will yield an empirical sequence of numbers. We postulate the existence of a function $F(x)$ (the properties of this function are given in §2.11) to serve as a mathematical model for $F_\infty(x)$. In some situations such as coin tossing, dice throwing, etc., a complete numerical specification of $F(x)$ can be proposed by combinatorial and other a priori considerations. In other situations of a more purely statistical nature it may be impossible to specify $F(x)$ beyond a particular functional form involving certain parameters.

In attempting to relate the behavior of the empirical cumulative distribution function $F_n(x)$ to the mathematical abstraction $F(x)$ one encounters at least two difficulties: One is common to all mathematical theories of physical (chemical, biological, sociological) phenomena, employing limits: the mathematical process of passing through an infinite number of steps is physically unrealizable, and is often impossible even as a "thought-experiment". For example, let the reader consider the notion of mass or charge density in the light of the fact that mass and charge are discrete. The other difficulty is peculiar to probability theory in that the theory does not assert that $\lim_{n\to\infty} F_n(x) = F(x)$, but that the approach* is in a sense defined within the framework of the theory itself: $F_n(x)$ converges <u>stochastically</u> to $F(x)$. Stochastic convergence is defined in §4.21.

Once $F(x)$ has been postulated, the mathematics begins and it consists of carrying out various mathematical manipulations on $F(x)$ corresponding to certain operations which can be performed on the sequence produced by the given randomizing process. The mathematics then becomes a method of making predictions of what will happen if certain operations are applied to the sequence. For example, $F(b)-F(a)$ is a prediction of the proportion of times, in a large number of trials, that the given process will yield numbers greater than a and less than or equal to b; $\int_{-\infty}^{+\infty} x\, dF(x)$ (taken in the Stieltjes sense, §2.5) is a prediction of the average of numbers obtained in a long series of repeated applications of the process; $F(x_1)\cdot F(x_2)$ is a prediction of the proportion of samples of pairs of numbers, out of a large number of such pairs, in which the first number is $\leq x_1$ and the second $\leq x_2$; $\iint_R dF(x_1)dF(x_2)$, where R is the region in the x_1x_2 plane for which $A\leq\frac{1}{2}(x_1+x_2)\leq B$, is a prediction of the proportion of samples of pairs of numbers, out of a large number of such pairs, in which the average of the sample pair lies between A and B. Many other examples could be given here but these will perhaps illustrate the nature of the correspondence between the mathematical operations performed on $F(x)$ (i. e. probability theory) and calculations based on the results of repeated applications of a given randomizing process. The degree of correspondence, i. e. validity of prediction, depends on the degree of randomness in the empirical sequence and on how well the function $F(x)$ has been chosen. That such predictions, correctly applied, have practical validity has been experimentally verified many times.

---

*See a study by V. I. Smirnoff, "Sur les ecarts de la courbe de distribution empirique", <u>Recueil Mathématique</u>, Moscow, vol. 6 (1939), pp. 25-26.

CHAPTER II

DISTRIBUTION FUNCTIONS

In this chapter we outline the basic probability theory necessary for the work of the course. The treatment is general, the study of important particular distributions being postponed to the next chapter.

2.1 Cumulative Distribution Functions

In the previous chapter we have introduced the notion of an empirical cumulative distribution function (c. d. f.) $F_n(x)$, and have indicated that it is an experimental fact that $F_n(x)$ appears to approach a limiting form $F_\infty(x)$ as n is increased. We now define a mathematical model F(x) for the intuitively apprehended $F_\infty(x)$ by laying down postulates for distribution functions. Henceforth the term cumulative distribution function (c. d. f.) will be used only in the sense defined below.

We shall find it convenient to use the following notations and definitions from point set theory: $P \in E$ signifies that the point P belongs to the set E. $E_1 \supset E_2$ is read "E_1 contains E_2". The sum (or union) of E_1 and E_2 is the totality of points P for which $P \in E_1$ or E_2; we shall denote it by $E_1 + E_2$. The product (or intersection) of E_1 and E_2 is the totality of points P for which $P \in$ both E_1 and E_2; we write it E_1E_2. E_1 and E_2 are said to be disjoint if they have no points in common. The difference $E_1 - E_2$ is the totality of points in E_1 not in E_2.

2.11 Univariate Case

A c. d. f. F(x) is defined by the following postulates:

1) If $x' < x''$, then $F(x'') - F(x') \geq 0$.

2) $F(-\infty) = 0$, $F(+\infty) = 1$.

The notation in (2) implies that the limits of F(x) exist as $x \longrightarrow -\infty$ or $+\infty$. Since (1) means that F(x) is monotone, it follows that F(x) has at most an enumerable number of discontinuities, and that the limits F(x+0), F(x-0) exist everywhere. The determination of the values of F(x) at its discontinuities is really not essential, but it will be convenient to fix them by

3) $F(x+0) = F(x)$.

It follows from (1) and (2) that $F(x)$ is non-negative.

The relation between probability statements about a random variable* X and its c. d. f. is determined by the following further postulates:

1') $\Pr(X \leqq x) = F(x)$.

The left member is read "the probability that $X \leqq x$." Let $E_1, E_2, \ldots\ldots$, be a finite or enumerable number of disjoint point sets on the x-axis:

2') $\Pr(X \epsilon E_1 + E_2 + \ldots) = \Pr(X \epsilon E_1) + \Pr(X \epsilon E_2) + \ldots$ This may be called the law of complete additivity, and may be used to determine the term on the left side of the equation, or any term on the right, when all the other probabilities entering the equation are known. For example, let I be the interval $x' < x \leqq x''$, I' be the interval $-\infty < x \leqq x'$, I" be the interval $-\infty < x \leqq x''$. Then

$$I'' = I' + I.$$

From (1')

$$\Pr(X \epsilon I') = F(x'), \qquad \Pr(X \epsilon I'') = F(x''),$$

and hence from (2') we may state the theorem

A) $\Pr(x' < X \leqq x'') = F(x'') - F(x')$.

In order to find the probability that X be equal to a given value x' take a sequence of points $a_1 < a_2 < a_3 < \ldots$ converging to x'. Let I be the interval $a_1 < x \leqq x'$, and I_j be the interval $a_j < x \leqq a_{j+1}$. Then

$$I = x' + \sum_{j=1}^{\infty} I_j.$$

Hence from (2'),

$$\Pr(X \epsilon I) = \Pr(X = x') + \sum_{j=1}^{\infty} \Pr(X \epsilon I_j),$$

and from theorem (A),

$$F(x') - F(a_1) = \Pr(X = x') + \sum_{j=1}^{\infty} [F(a_{j+1}) - F(a_j)].$$

---

*In this chapter it is convenient to denote a random variable by a capital letter, X, etc., and the corresponding independent variable in the distribution function by the corresponding lower case letter, x, etc. In later chapters we will drop this convention when there is no danger of confusion.

Now

$$\sum_{j=1}^{\infty}[F(a_{j+1})-F(a_j)] = \lim_{n\to\infty}\sum_{j=1}^{n}[F(a_{j+1})-F(a_j)] = \lim_{n\to\infty}[F(a_{n+1})-F(a_1)] = F(x'-0) - F(a_1).$$

Hence we have the theorem

$$\text{B)}\quad \Pr(X=x) = F(x) - F(x-0).$$

In a similar manner one may derive the following theorems:

$$\text{C)}\quad \Pr(x'<X<x'') = F(x''-0) - F(x'),$$
$$\Pr(x'\leq X\leq x'') = F(x'') - F(x'-0),$$
$$\Pr(x'\leq X<x'') = F(x''-0) - F(x'-0).$$

$$\text{D)}\quad 0 \leq \Pr(X\in E) \leq 1 \quad \text{(for any set E for which the middle member is defined).}$$

$$\text{E)}\quad \Pr(-\infty < X < +\infty) = 1.$$

Let $E_1, E_2, \ldots,$ be sets which are not necessarily disjoint, then

$$\text{F)}\quad \Pr(X\in E_1+E_2) = \Pr(X\in E_1) + \Pr(X\in E_2) - \Pr(X\in E_1E_2),$$
$$\Pr(X\in E_1+E_2+E_3) = \sum_{j=1}^{3}\Pr(X\in E_j) - \sum_{i<j}\Pr(X\in E_iE_j) + \Pr(X\in E_1E_2E_3), \text{ etc.}$$

We now characterize two important classes of c. d. f.'s:

(1) Suppose that F(x) increases only by jumps, -- more precisely, suppose a finite or an enumerable set of points $x_1, x_2, \ldots,$ and corresponding positive numbers $p_1, p_2, \ldots, \sum p_i=1$, such that $F(x) = \sum p_j$ summed over all j for which $x_j\leq x$. We shall call this the <u>discrete</u>* case. It may be shown from the theorem (B) that in this case $\Pr(X=x_i) = p_i$, while for any point $x' \neq$ any x_i, $\Pr(X=x') = 0$.

If the number of x_i is finite, or more generally, if the x_i have no cluster points except $\pm\infty$, then the graph of F(x) in this case is a step-function made up of horizontal lines as shown in (a) of Figure 1. The jump at $x = x_i$ is equal to p_i, the

---

*It should be noted that an empirical c. d. f. $F_n(x)$ of an observation variable X has properties (1), (2), (3) of a c. d. f. (discrete case). $F_n(x)$ does not have properties (1') and (2'), although it has analogous properties. That is, corresponding to (1') we would have $\text{Prop}(X\leq x) = F_n(x)$ (proportion of values of $X\leq x$ is $F_n(x)$) and for (2') we would have $\text{Prop}(X\in E_1+E_2+\ldots+E_k) = \text{Prop}(X\in E_1) + \text{Prop}(X\in E_2) + \ldots + \text{Prop}(X\in E_k)$. Thus, in the case of $F_n(x)$, P_i would be the <u>proportion of cases</u> among the n values of X, in which the observation variable = x_i and not <u>probability</u> that $X = x_i$.

probability that $X = x_1$.

(ii) Another case is characterized by the existence of a function $f(x) \geq 0$ such that

$$F(x) = \int_{-\infty}^{x} f(\xi)\, d\xi.$$

This is really a necessary and sufficient condition for the absolute continuity of $F(x)$, but instead of calling this the absolutely continuous case, we shall refer to it merely as the continuous case. The graph of $F(x)$ in this case is continuous as shown in (b) of Figure 1. We shall call $f(x)$ the probability density function of the random variable X. The reader may show that in this case

$$\Pr(x' \leqq X \leqq x'') = \int_{x'}^{x''} f(\xi)\, d\xi,$$

and that the statement remains valid if one or both of the equality signs inside the parentheses on the left are deleted. If $f(x)$ is continuous for $x' < x < x''$,

$$\Pr(x' \leqq X \leqq x'') = (x''-x')\, f(x_1), \text{ where } x' < x_1 < x'',$$

and if $f(x)$ is continuous at x_0,

$$\Pr(x_0 \leqq X \leqq x_0+dx) = f(x_0)\, dx,$$

except for infinitesimals of higher order. The infinitesimal $f(x)dx$ is sometimes called the probability element of X.

The discrete and continuous cases thus defined obviously do not cover all univariate c. d. f.'s, but we shall confine ourselves to these in the present course.

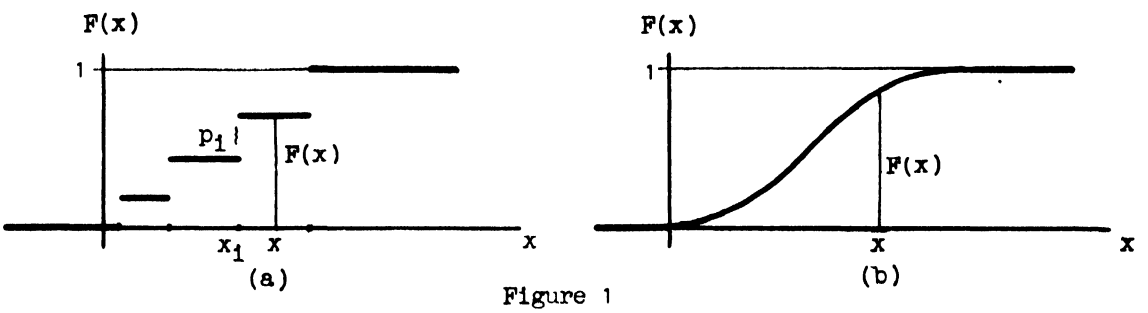

Figure 1

2.12 Bivariate Case

Let J be a rectangle in the x_1, x_2 plane, $x_1' < x_1 \leqq x_1''$, $x_2' < x_2 \leqq x_2''$. Denote by $\Delta_J^2 F(x_1, x_2)$ the second difference

$$\Delta_J^2 F(x_1,x_2) = F(x_1'',x_2'') + F(x_1',x_2') - F(x_1',x_2'') - F(x_1'',x_2').$$

Then a c. d. f. $F(x_1,x_2)$ is subjected to the following postulates:

$$1)\quad \Delta_J^2 F(x_1,x_2) \geq 0.$$

$$2)\quad F(-\infty,x_2) = F(x_1,-\infty) = 0, \qquad F(+\infty,+\infty) = 1.$$

By letting $x_1' \longrightarrow -\infty$ in (1), we get with the aid of (2),

$$F(x_1,x_2'') - F(x_1,x_2') \geq 0 \text{ if } x_2'' > x_2',$$

and similarly

$$F(x_1'',x_2) - F(x_1',x_2) \geq 0 \text{ if } x_1'' > x_1',$$

so that $F(x_1,x_2)$ is monotonic in each variable separately. Hence the limits $F(x_1\pm 0,x_2)$, $F(x_1,x_2\pm 0)$ exist everywhere. It can be shown that $F(x_1,x_2)$ is discontinuous in x_1 at worst on an enumerable number of lines x_1 = constant, and similarly for x_2. If we let $x_1' \longrightarrow -\infty$ and $x_2' \longrightarrow -\infty$ in (1), we get $F(x_1,x_2) \geq 0$ because of (2). The values of $F(x_1,x_2)$ at its discontinuities are fixed by

$$3)\quad F(x_1,x_2) = F(x_1+0,x_2) = F(x_1,x_2+0).$$

The tieup of probability statements about a vector random variable X_1,X_2 with two components with its c. d. f. is determined by the following further postulates:

$$1')\quad \Pr(X_1 \leq x_1, X_2 \leq x_2) = F(x_1,x_2).$$

Let E_1, E_2,...., be disjoint sets, then

$$2')\quad \Pr(X_1,X_2 \in E_1+E_2+\ldots) = \Pr(X_1,X_2 \in E_1) + \Pr(X_1,X_2 \in E_2) + \ldots$$

By methods of §2.11 the reader may verify the following theorems:

$$A)\quad \Pr(X_1,X_2 \in J) = \Delta_J^2 F(x_1,x_2),$$

where J and Δ_J^2 are defined above.

$$B)\quad \Pr(x_1' < X_1 \leq x_1'', X_2 = x_2) = F(x_1'',x_2) + F(x_1',x_2-0) - F(x_1',x_2) - F(x_1'',x_2-0).$$

$$C)\quad \Pr(X_1 = x_1, X_2 = x_2) = F(x_1,x_2) + F(x_1-0,x_2-0) - F(x_1-0,x_2) - F(x_1,x_2-0).$$

It can be shown by methods beyond the level of this course that from the postulates (1'), (2') the probability that $X_1,X_2 \in E$ is determined for a very general class of regions

called Borel-measurable* regions.

D) $0 \leqq \Pr(X_1, X_2 \in E) \leqq 1$.

E) $\Pr(-\infty < X_1 < +\infty, -\infty < X_2 < +\infty) = 1$.

For sets E_1, E_2,...., not necessarily disjoint,

F) Theorem (F) of §2.11 is valid

With a bivariate distribution function we shall be mainly interested in the discrete case and the continuous case, and occasionally a mixed case, all defined below. We remark again that these categories are not exhaustive.

i) The discrete** case is characterized by the existence of a finite or enumerable set of points (x_{1i}, x_{2i}), $i=1,2,\ldots$, and associated positive numbers p_i (probabilities) $\sum p_i = 1$, such that $F(x_1, x_2) = \sum p_j$ summed for all j for which $x_{1j} \leqq x_1$ and $x_{2j} \leqq x_2$. From theorem (C) it follows that $\Pr(X_1 = x_{1i}, X_2 = x_{2i}) = p_i$, and for any point (x_1', x_2') not in the set (x_{1i}, x_{2i}), $\Pr(X_1 = x_1', X_2 = x_2') = 0$.

ii) By the continuous case (see remarks in §2.11 about absolute continuity) we shall understand that in which there exists a function $f(x_1, x_2) \geqq 0$ such that

$$\text{(a)} \qquad F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(\xi_1, \xi_2)\, d\xi_1 d\xi_2.$$

We may show that

$$\Pr(X_1, X_2 \in J) = \iint_J f(x_1, x_2)\, dx_1 dx_2,$$

---

*In k-dimensional space a Borel-measurable region (or a Borel set) is one that is obtainable from half-open intervals or cells, $x_i' < x_i \leqq x_i''$, $i = 1, 2, \ldots, k$, by taking a finite or enumerable number of sums, differences, and products of such cells. A function $f(x)$ is Borel-measurable if the set of values of x for which $a < f(x) \leqq b$ is a Borel set, where a and b are any two real numbers. A Borel-measurable function of two or more variables is similarly defined.

**As in the case of one variable, it should be observed that an empirical c. d. f. $F_n(x_1, x_2)$ of two observation variables X_1, X_2 has properties (1), (2), (3) of a c. d. f. for two random variables (discrete case). But, in (1') and (2') one would use the term "proportion of cases" instead of the term "probability of". The p_i associated with the isolated points (x_{1i}, x_{2i}) would be called the proportion of cases for which $X_1 = x_{1i}$, $X_2 = x_{2i}$, instead of the probability that $X_1 = x_{1i}$, $X_2 = x_{2i}$. The number of such points would be $\leqq n$, the number of observed pairs of values of X_1, X_2.

This comparison of an empirical c. d. f. and the case of discrete variables extends at once to the case of k variables discussed in §2.13.

and that the result is not invalidated if J is closed by the addition of its boundaries. From this it follows that, except for infinitesimals of higher order,

$$\Pr(x_1 \leqq X_1 \leqq x_1+dx_1, x_2 \leqq X_2 \leqq x_2+dx_2) = f(x_1,x_2)\, dx_1 dx_2.$$

$f(x_1,x_2)$ and $f(x_1,x_2)dx_1dx_2$ are called respectively the p. d. f.* and the probability element of the random variables X_1, X_2.

iii) The mixed** case (X_1 continuous, X_2 discrete) is said to obtain if there exists a finite or enumerable set of lines $x_2 = x_{2i}$, $i = 1, 2, \ldots$, associated positive numbers p_{2i}, $\sum_I p_{2i} = 1$, and a non-negative function of x_1 and x_2 defined for all x_1 and for $x_2 = x_{2i}$, $i = 1, 2, \ldots$, which function we shall write as $f(x_1|x_{2i})$, such that

$$\int_{-\infty}^{+\infty} f(\xi_1|x_{2i})d\xi_1 = 1, \qquad i = 1, 2, \ldots,$$

and

$$F(x_1,x_2) = \sum p_{2j} \int_{-\infty}^{x_1} f(\xi_1|x_{2j})d\xi_1, \text{ summed over all } j \text{ for which } x_{2j} \leqq x_2.$$

In the mixed case p_{2i} is the probability that the random point X_1, X_2 will fall on the line $x_2 = x_{2i}$, and $f(x_1|x_2)dx_1$ is the probability (to within terms of order dx_1) that $x_1 < X_1 < x_1+dx_1$ if the random point falls on the line $x_2 = x_{2i}$.

It may be shown from our postulates that for any (B-meas.) region E in the x_1x_2-plane we get in the three cases

$$\Pr(X_1,X_2 \in E) = \begin{cases} \text{i)} & \sum p_j \text{ summed over all } j \text{ such that } x_{1j}, x_{2j} \in E, \\ \text{ii)} & \iint_E f(x_1,x_2)dx_1dx_2, \\ \text{iii)} & \sum_{\text{all } i} p_{2i} \int_{E_{2i}} f(x_1|x_{2i})dx_1, \end{cases}$$

where E_{2i} is the projection on the x_1 axis of the part of the line $x_2=x_{2i}$ lying in E. (If the line does not intersect E, the corresponding integral is zero.)

By means of the Stieltjes integral (§2.5) these three cases may be brought under the single expression $\Pr(X_1,X_2 \in E) = \int_E dF(x_1,x_2)$, which includes indeed the most general case.

2.13 k-Variate Case

A k-variate c. d. f. $F(x_1,x_2,\ldots,x_k)$ must satisfy the following three postulates: Let J be the k-dimensional cell $x_i' < x_i \leqq x_i''$, $i=1,2,\ldots,k$, and define the k-th difference

---

*probability density function

**The reader will understand this case better if he rereads this description after having mastered §2.4.

$$\Delta_J^k F(x_1,x_2,\ldots,x_k) = \Delta_1\Delta_2\ldots\Delta_{k-1}\Delta_k F(x_1,x_2,\ldots,x_k),$$

where the operators Δ_1 are applied successively and denote

$$\Delta_1 F(x_1,x_2,\ldots,x_k) = F(x_1,\ldots,x_{i-1},x_i'',x_{i+1},\ldots,x_k) - F(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_k).$$

1) $\Delta_J^k F(x_1,x_2,\ldots,x_k) \geq 0.$

2) $F(-\infty,x_2,\ldots,x_k) = F(x_1,-\infty,x_3,\ldots,x_k) = \ldots$
$= F(x_1,\ldots,x_{k-1},-\infty) = 0, \quad F(+\infty,+\infty,\ldots,+\infty) = 1.$

3) $F(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_k) = F(x_1,\ldots,x_{i-1},x_i+0,x_{i+1},\ldots,x_k).$

As in the bivariate case it can be shown from (1) and (2) that F is monotonic in each variable separately and that F is monotonic (in the sense of (1)) in any set of variables if the remainder are held fixed.

A random vector variable $X = (X_1,X_2,\ldots,X_k)$ is said to have the c. d. f. $F(x_1,x_2,\ldots,x_k)$, --or the random variables $X_1,X_2,\ldots,X_k$ are said to be jointly distributed with the c. d. f. --if furthermore

1') $\Pr(X_1\leq x_1, X_2\leq x_2,\ldots,X_k\leq x_k) = F(x_1,x_2,\ldots,x_k).$

If $E_1,E_2,\ldots,$ are a finite or enumerable number of disjoint sets,

2') $\Pr(X\in E_1+E_2+\ldots) = \Pr(X\in E_1) + \Pr(X\in E_2) + \ldots$

By the methods used before we may now generalize the theorems (A) to (F) of §§2.11 and 2.12.

The discrete case and the continuous are defined by obvious generalization of §2.12, and it is evident how mixed cases of various orders would now be defined.

2.2 Marginal Distributions

Suppose the joint c. d. f. of the random variables X_1,X_2 is $F(x_1,x_2)$, and consider the probability that $X_1 \leq x_1$, without any condition on X_2:

$$\Pr(X_1\leq x_1) = \Pr(X_1\leq x_1, X_2<+\infty) = F(x_1,+\infty).$$

This is called the marginal distribution of X_1. We note that it is a bona fide distribution function as defined in §2.11, in fact, it is the univariate c. d. f. of X_1. Similarly, we define $F(+\infty,x_2)$ as the marginal distribution of X_2.

For the discrete case defined in §2.12 we then have

$$F(x_1,+\infty) = \sum p_j \text{ summed for all } j \text{ such that } x_{1j} \leq x_1.$$

For the continuous case,

$$\text{(a)} \qquad F(x_1,+\infty) = \int_{-\infty}^{x_1}\int_{-\infty}^{+\infty} f(\xi_1,\xi_2)d\xi_1 d\xi_2 = \int_{-\infty}^{x_1} f_1(\xi_1)d\xi_1,$$

where

$$f_1(x_1) = \int_{-\infty}^{+\infty} f(x_1,\xi)d\xi.$$

$f_1(x_1)$ may be called the marginal p. d. f. of X_1.

In the trivariate case we get besides the marginal distribution of each random variable separately, for example,

$$F(x_1,+\infty,+\infty) = \Pr(X_1 \leqq x_1),$$

also marginal distributions of pairs of random variables, for example,

$$F(x_1,x_2,+\infty) = \Pr(X_1 \leqq x_1, X_2 \leqq x_2).$$

For a k-variate distribution one likewise defines marginal distributions of the random variables taken one at a time, in pairs,..., k-1 at a time. We note that all these marginal distributions satisfy the postulates (1), (2), (3), (1'), (2') for a c. d. f.

2.3 Statistical Independence

If $F(x_1,x_2)$ is the c. d. f. of X_1,X_2, then from §2.2,

$$F_1(x_1) = F(x_1,+\infty), \quad F_2(x_2) = F(+\infty,x_2)$$

are the marginal distributions of X_1 and X_2, respectively. We say that the random variables X_1,X_2 are independent in the probability sense, or statistically independent, if

$$\text{(a)} \qquad F(x_1,x_2) = F_1(x_1)\,F_2(x_2).$$

It is easily seen that a necessary and sufficient condition for the statistical independence of X_1 and X_2 is that their joint c. d. f. factor into a function of x_1 alone times a function of x_2 alone, i. e.,

$$F(x_1,x_2) = G(x_1)\,H(x_2).$$

In order to see the probability implications of statistical independence, consider any two intervals I_1 and I_2 on the x_1 and x_2-axes, respectively,

$$I_1: \; x_1' < x_1 \leqq x_1'',$$
$$I_2: \; x_2' < x_2 \leqq x_2'',$$

and let J be the rectangle of points (x_1,x_2) satisfying both these inequalities. Then

(b) $$\Pr(X_1,X_2\epsilon J) = \Pr(X_1\epsilon I_1)\cdot\Pr(X_2\epsilon I_2).$$

For, by hypothesis, we have (a); hence

$$\Pr(X_1,X_2\epsilon J) = \Delta_J^2 F(x_1,x_2) = F_1(x_1'')F_2(x_2'') + F_1(x_1')F_2(x_2') - F_1(x_1')F_2(x_2'') - F_1(x_1'')F_2(x_2').$$

After factoring the right member we easily get (b).

By the same method, and with the aid of Theorem (B) of §2.11 and Theorem (C) of §2.12 we get that if X_1 and X_2 are statistically independent, then

$$\Pr(X_1=x_1,X_2=x_2) = \Pr(X_1=x_1)\cdot\Pr(X_2=x_2).$$

This is of importance for the discrete case. For the continuous case we may state the following result: If $f(x_1,x_2)$ is the joint p. d. f. of X_1,X_2, if $f_j(x_j)$ is the marginal p. d. f. of X_j, $j = 1,2$, and if X_1,X_2 are statistically independent, then

$$f(x_1,x_2) = f_1(x_1)\, f_2\,(x_2)$$

wherever $f(x_1,x_2)$ is continuous. At the points of continuity, we have from equation (a) of §2.12,

$$\begin{aligned} f(x_1,x_2) &= \frac{\partial^2}{\partial x_1\partial x_2}F(x_1,x_2) \\ &= \frac{\partial^2}{\partial x_1\partial x_2}\{F_1(x_1)F_2(x_2)\} \\ &= \frac{\partial F_1(x_1)}{\partial x_1}\cdot\frac{\partial F_2(x_2)}{\partial x_2} \\ &= f_1(x_1)f_2(x_2), \end{aligned}$$

the last step following from (a) of §2.2.

k random variables are said to be <u>mutually</u> (statistically) <u>independent</u> if their joint c. d. f. is of the form

$$F(x_1,x_2,\ldots,x_k) = F_1(x_1)F_2(x_2)\ldots F_k(x_k),$$

where $F_j(x_j)$ is the marginal distribution of X_j. Two random vector variables $X_i = (X_{i1},X_{i2},\ldots,X_{ik_i})$, $i = 1,2$, are called <u>statistically independent</u> if the joint c. d. f. of the $k_1 + k_2$ components is the product of the marginal distributions of X_1 and X_2:

$$F(x_{11},\ldots,x_{1k_1};x_{21},\ldots,x_{2k_2}) = F(x_{11},\ldots,x_{1k_1};+\infty,\ldots,+\infty)\cdot F(+\infty,\ldots,+\infty;x_{21},\ldots,x_{2k_2}).$$

The definition of the statistical independence of n vector random variables is made as

the obvious generalization.

The concept of statistical independence is fundamental in sampling theory (§4.1). n random variables are said to constitute a random sample from a population (§4.1) with c. d. f. $F(x)$ if their joint c. d. f. is $F(x_1)F(x_2)\ldots F(x_n)$. If the population distribution is k-variate with c. d. f. $F(x_1,x_2,\ldots,x_k)$, then the n vector variables $X_i = (X_{i1},X_{i2},\ldots,X_{ik})$, $i = 1, 2,\ldots,n$, are said to be a random sample if the joint c. d. f. of the set $|X_{i\alpha}|$ is

$$\prod_{i=1}^{n} F(x_{i1},x_{i2},\ldots,x_{ik}).$$

2.4 Conditional Probability

Let X be a random variable, and let R be any (Borel) set of points on the x-axis. Let E be any (Borel) subset of R. If $\Pr(X\epsilon R) \neq 0$, we define the conditional probability $\Pr(X\epsilon E|X\epsilon R)$, read "the probability that X is in E, given that X is in R", as

$$\text{(a)} \qquad \Pr(X\epsilon E|X\epsilon R) = \frac{\Pr(X\epsilon E)}{\Pr(X\epsilon R)}.$$

The definition (a) extends immediately to any finite number of random variables. For example for two random variables X_1, X_2, R would represent a (Borel) set in the x_1x_2 plane and E would be a subset of R.

Of particular interest is the case in which R is a set in the x_1x_2 plane for which $X_1\epsilon E_1$ where E_1 is any (Borel) set in the domain of X_1, and E is the product or inter section set between R and a similar set for which $X_2\epsilon E_2$, where E_2 is any (Borel) set in the domain of X_2. Here we may write $E = E_1E_2$. The simplest case is that in which E_1 is an interval $x_1' < x_1 \leq x_1''$ and E_2 is an interval $x_2' < x_2 \leq x_2''$. Then R is the horizontal strip $x_2' < x_2 \leq x_2''$, and E is the rectangle for which $x_1' < x_1 \leq x_1''$ and $x_2' < x_2 \leq x_2''$. In the present case, expression (a) may be written in the form

$$\text{(b)} \qquad \Pr(X_1\epsilon E_1|X_2\epsilon E_2) = \frac{\Pr(X_1,X_2\epsilon E)}{\Pr(X_2\epsilon E_2)}.$$

Because of symmetry, we may also write

$$\Pr(X_2\epsilon E_2|X_1\epsilon E_1) = \frac{\Pr(X_1,X_2\epsilon E)}{\Pr(X_1\epsilon E_1)}.$$

In a similar manner we may write for the case of three variables

$$\Pr(X_3\epsilon E_3|X_1,X_2\epsilon E_1E_2) = \frac{\Pr(X_1,X_2,X_3\epsilon E_1E_2E_3)}{\Pr(X_1,X_2\epsilon E_1E_2)},$$

and so on for any number of variables. The relation (b) may of course be expressed in

terms of distribution functions. In particular, if X_1, X_2 have a bivariate p. d. f. $f(x_1,x_2)$, and E_1 is the set

(c) $$x_1' \leq x_1 \leq x_1''$$

on the x_1-axis, and E_2 is the set

(d) $$x_2' \leq x_2 \leq x_2' + h$$

on the x_2-axis, then E is the rectangle in the x_1x_2-plane defined by (c) and (d). Equation (b) becomes

(e) $$\Pr(x_1' \leq X_1 \leq x_1'' \mid x_2' \leq X_2 \leq x_2'+h) = \frac{\int_{x_2'}^{x_2'+h} \int_{x_1'}^{x_1''} f(x_1,x_2)\,dx_1\,dx_2}{\int_{x_2'}^{x_2'+h} f_2(x_2)\,dx_2}$$

if the denominator does not vanish. If $f(x_1,x_2)$ is continuous in the rectangle E, the denominator may be written

$$h\, f_2(\xi_2), \text{ where } x_2' < \xi_2 < x_2' + h,$$

and the numerator,

$$\int_{x_1'}^{x_1''} h\, f(x_1,\eta_2(x_1))\,dx_1, \text{ where } x_2' < \eta_2(x_1) < x_2' + h.$$

(e) may then be written

(f) $$\int_{x_1'}^{x_1''} [f(x_1,\eta_2(x_1))/f_2(\xi_2)]\,dx_1.$$

We note that the integrand, for fixed x_2' and h, has the properties of a univariate p. d. f. We next assume that $f_2(x_2') \neq 0$. Noting that $\Pr(x_1' \leq X_1 \leq x_1'' \mid X_2 = x_2')$ is not defined by (b), we now define it as the limit of (e) as $h \longrightarrow 0$. The continuity we have already assumed is sufficient to justify our taking limits under the integral sign in (f); the result is

$$\Pr(x_1' \leq X_1 \leq x_1'' \mid X_2 = x_2) = \int_{x_1'}^{x_1''} f(x_1 \mid x_2)\,dx_1,$$

where

(g) $$f(x_1 \mid x_2) = f(x_1,x_2)/f_2(x_2).$$

For fixed x_2, $f(x_1|x_2)$ again has the properties of a univariate p. d. f.; it may be called the conditional p. d. f. of x_1, given x_2. We note that if X_1 and X_2 are statistically independent, $f(x_1|x_2) = f_1(x_1)$.

Likewise, if random variables $X_{11},\ldots,X_{1k_1};X_{21},\ldots,X_{2k_2}$ have a joint p. d. f. $f(x_{11},\ldots,x_{1k_1};x_{21},\ldots,x_{2k_2})$ we define the conditional p. d. f.

$$\text{(h)}\quad f(x_{11},\ldots,x_{1k_1}|x_{21},\ldots,x_{2k_2}) = \frac{f(x_{11},\ldots,x_{1k_1};x_{21},\ldots,x_{2k_2})}{\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(x_{11},\ldots,x_{1k_1};x_{21},\ldots,x_{2k_2})\,dx_{11}\cdots dx_{1k_1}},$$

if the denominator is not zero.

2.5 The Stieltjes Integral

An important tool in mathematical statistics, which often permits a common treatment of the discrete and continuous cases (and indeed the most general case), is the Stieltjes integral.

2.51 Univariate Case

We begin by defining the Stieltjes integral over a finite half-open interval $a < x \leqq b$: Suppose we have two functions, $\phi(x)$ continuous for $a \leqq x \leqq b$, and $F(x)$ monotone for $a \leqq x \leqq b$. We subdivide (a,b) into subintervals $I_j:(x_{j-1},x_j)$ by means of points $x_0 = a < x_1 < x_2 < \ldots < x_m = b$. In each interval we pick an arbitrary point $\xi_j \in I_j$. Denote by $\Delta_j F(x)$ the difference $F(x_j) - F(x_{j-1})$, and form the sum

$$S = \sum_{j=1}^{m} \phi(\xi_j)\,\Delta_j F(x).$$

If U_j is the maximum of $\phi(x)$ in I_j, and L_j, the minimum, then

$$\text{(a)}\qquad \begin{aligned} L_j &\leqq \phi(\xi_j) \leqq U_j, \\ S_L &\leqq S \leqq S_U, \end{aligned}$$

where

$$S_L = \sum_{j=1}^{m} L_j \Delta_j F(x),$$

$$S_U = \sum_{j=1}^{m} U_j \Delta_j F(x).$$

Let $\epsilon = \max_j (U_j - L_j)$. Then

$$0 \leqq S_U - S_L = \sum_{j=1}^{m} (U_j - L_j)\Delta_j F(x) \leqq \epsilon \sum_{j=1}^{m} \Delta_j F(x) = \epsilon[F(b)-F(a)].$$

Hence if the intervals I_j are further subdivided, and this process is continued in such a way that the norm of the subdivision, $\delta = \max_j(x_j - x_{j-1})$, approaches zero, then since $\phi(x)$ is uniformly continuous on (a,b), $\epsilon \longrightarrow 0$, and hence

$$S_U - S_L \longrightarrow 0.$$

It is easily seen that S_L is non-decreasing, and S_U, non-increasing, as the subdivision is made finer, and hence from (a), S approaches a limit. Since S_L and S_U are independent of the choice of the arbitrary point ξ_j in I_j, therefore from (a), $\lim_{\delta \to 0} S$ is likewise independent of this choice. Furthermore, $\lim_{\delta \to 0} S$ may be shown to be independent of the method of subdivision. We call this limit the Stieltjes integral of $\phi(x)$ with respect to $F(x)$ over the range $a < x \leq b$ and denote it by

$$\int_a^b \phi(x)dF(x) = \lim_{\delta \to 0} S.$$

Let us examine further the significance of the Stieltjes integral when $F(x)$ is a c. d. f. in the discrete or continuous cases: Suppose that $F(x)$ is a discrete c. d. f. with only a finite number n of jumps of amount p_k at the points a_k in the interval (a,b). We may assume that the points are ordered,

$$(b) \qquad a < a_1 < a_2 < \ldots < a_n \leq b.$$

Since the points a_k are isolated, eventually for any mode of continued subdivision, each interval I_j will contain not more than one point a_k in its interior or as right end point. If I_j contains a_k, that is if $x_{j-1} < a_k \leq x_j$, denote it by I_k', and call the arbitrary point ξ_j in this interval, ξ_k'. Then

$$\Delta_j F(x) = \begin{cases} p_k \text{ if } I_j = I_k', \\ 0 \text{ if } I_j \text{ is not an } I_k'. \end{cases}$$

Hence

$$S = \sum_{k=1}^{n} \phi(\xi_k')\, p_k.$$

Now as the norm $\delta \longrightarrow 0$, $\xi_k' \longrightarrow a_k$, $\phi(\xi_k') \longrightarrow \phi(a_k)$, and thus

$$(c) \qquad \int_a^b \phi(x)dF(x) = \sum_{k=1}^{n} \phi(a_k)\, p_k.$$

It will be noted that the continuity of $\phi(x)$ at the points a_k is essential. The result (c) may be shown to remain valid in the case where there is an infinite number of points of discontinuity of $F(x)$ in (a,b).

In the continuous case at points of continuity of the p. d. f. $f(x)$ we have

$$dF(x)/dx = f(x), \qquad dF(x) = f(x)dx,$$

and hence we might write heuristically

$$\text{(d)} \qquad \int_a^b \phi(x)dF(x) = \int_a^b \phi(x)f(x)dx.$$

The relation (d) may be proved as follows: We first assume that $f(x)$ is continuous on (a,b). Then in each interval I_j we pick as ξ_j the point for which

$$\Delta_j F(x) = F'(\xi_j)(x_j - x_{j-1}).$$

The existence of such a point is guaranteed by the mean value theorem. Then

$$S = \sum_{j=1}^{m} \phi(\xi_j)f(\xi_j)(x_j - x_{j-1}).$$

But by the so-called fundamental theorem of the calculus (actually, the definition of the ordinary definite integral), the limit of this sum as the norm approaches zero is the right member of (d). The proof can be extended to the case where $f(x)$ has discontinuities on (a,b).

We shall have need of the Stieltjes integral over an infinite interval. We define it as

$$\text{(e)} \qquad \int_{-\infty}^{+\infty} \phi(x)dF(x) = \lim_{\substack{a \to -\infty \\ b \to +\infty}} \int_a^b \phi(x)dF(x)$$

if and only if the limit exists as $a \to -\infty$, and $b \to +\infty$, <u>independently</u>. In more advanced work it is sometimes convenient to consider

$$\text{(f)} \qquad \lim_{T \to +\infty} \int_{-T}^{+T} \phi(x)dF(x).$$

This limit of course exists whenever (e) does, but the converse is false. (f) is called the <u>Cauchy principal value</u> of the infinite integral. Unless the contrary is explicitly stated, we shall always understand that the infinite integral connotes (e).

An intuitive explanation of the meaning of the Stieltjes integral will be given in §2.53, where we shall also indicate how the Stieltjes integral may be generalized over any range which is a Borel set E. In the univariate case, the various expressions for $\Pr(X \in E)$ introduced in §2.11 may then all be summarized under

$$\Pr(X \in E) = \int_E dF(x).$$

2.52 Bivariate Case

We limit our definition to the case where $F(x_1,x_2)$ is a c. d. f. as defined in §2.12. Let J be the half-open cell

(a) $$J: \quad a_1 < x_1 \leq b_1, \; a_2 < x_2 \leq b_2.$$

We assume $\phi(x_1,x_2)$ is continuous on J (boundaries included). By means of lines parallel to the axes, subdivide J into rectangles J_j, $j = 1,2,\ldots,m$. Let the norm δ of the subdivision be the maximum of the lengths of the diagonals of J_j. In each cell J_j pick a point (ξ_{1j},ξ_{2j}). Define $\Delta_j^2F(x_1,x_2)$, the second difference of $F(x_1,x_2)$ over the j-th cell, as in §2.12, and form the sum

$$S = \sum_{j=1}^{m} \phi(\xi_{1j},\xi_{2j})\Delta_j^2F(x_1,x_2).$$

By considering the upper and lower sums S_U and S_L, defined as in §2.51, we find again that $\lim_{\delta\to 0} S$ exists, and define it to be the Stieltjes integral of ϕ with respect to F over J:

(b) $$\int_J \phi(x_1,x_2)dF(x_1,x_2) = \lim_{\delta\to 0} S.$$

The remarks in §2.51 regarding the independence of (b) of the choice of (ξ_{1j},ξ_{2j}) and of the mode of subdivision remain valid.

As in §2.51 it may be shown that in the discrete case

$$\int_J \phi(x_1,x_2)dF(x_1,x_2) = \sum_{k=1}^{n} \phi(x_{1k},x_{2k})\, p_k,$$

where (x_{1k},x_{2k}) are the points in J (excluding the left and lower boundaries) where the probabilities are p_k (see §2.12). In the continuous case we may derive

$$\int_J \phi(x_1,x_2)dF(x_1,x_2) = \int_{a_2}^{b_2}\int_{a_1}^{b_1} \phi(x_1,x_2)f(x_1,x_2)dx_1dx_2.$$

In the mixed case defined in §2.12, and in the notation employed there, it may be shown that

$$\int_J \phi(x_1,x_2)dF(x_1,x_2) = \sum p_{2i}\int_{a_1}^{b_1} \phi(x_1,x_{2i})f(x_1|x_{2i})dx_1, \text{ summed for all i such that } a_2 < x_{2i} \leq b_2.$$

Denote by R_2 the entire x_1x_2-space. We say that the improper integral

$$\int_{R_2} \phi(x_1,x_2)dF(x_1,x_2)$$

exists if and only if the limit

$$\lim_{\substack{a_1 \to -\infty \\ b_1 \to +\infty}} \int_J \phi(x_1,x_2)dF(x_1,x_2)$$

exists, where J,a_1,b_1 are related by (a), as a_1,a_2,b_1,b_2 <u>independently</u> become infinite (with the signs indicated).

A generalization of the Stieltjes integral to regions more general than rectangles will be given in §2.53.

<u>2.53 k-Variate Case</u>

We first define the Stieltjes integral over a half-open cell,

(a) $$J: \quad a_i < x_i \leq b_i, \quad i = 1,2,\ldots,k.$$

We assume that $F(x_1,x_2,\ldots,x_k)$ is a k-variate c. d. f. as defined §2.13, and that $\phi(x_1,x_2,\ldots,x_k)$ is continuous in J (and on its boundaries). By means of hyperplanes x_i = constant, $i = 1,2,\ldots,k$, we subdivide J into cells J_j, $j = 1,2,\ldots,m$. Let δ be the length of the longest of the diagonals of the cells J_j. Define $\Delta_j^k F$. the k-th difference of F over the cell J_j as in §2.13, and form the sum

$$S = \sum_{j=1}^{m} \phi(\xi_{1j},\ldots,\xi_{kj})\Delta_j^k F,$$

where $(\xi_{1j},\ldots,\xi_{kj})$ is an arbitrary point in J_j. Under the hypotheses we have made, S converges to a limit independent of the choice of $(\xi_{1j},\ldots,\xi_{kj})$ and of the mode of subdivision, as $\delta \to 0$. We define

$$\int_J \phi(x_1,\ldots,x_k)dF(x_1,\ldots,x_k) = \lim_{\delta \to 0} S.$$

Let R_k be the entire $x_1x_2\ldots x_k$-space. The Stieltjes integral of ϕ with respect to F over R_k is defined as in §2.52.

Next, let us define the integral over a region K which is the sum of a finite or enumerable number of half-open cells J_i, $i = 1,2,\ldots,$

$$\int_K \phi dF = \sum_I \int_{J_i} \phi dF.$$

To define the integral over any (B-measurable) region E in R_k we cover E with a region of the type K just considered, and then take as the integral over E the greatest lower bound of the integral over K for all possible K containing E:

$$\int_E \phi dF = \text{g.l.b.}_{K \supset E} \int_K \phi dF.$$

In terms of our general definition of the Stieltjes integral we see that

$$\int_a^b \phi(x)dF(x) = \int_I \phi(x)dF(x)$$

only if I is the half-open interval $a < x \leq b$. For the closed interval we would have to add $\phi(a)[F(a)-F(a-0)] = \phi(a)Pr(X=a)$ to the left member; for the open interval, subtract $\phi(b)[F(b)-F(b-0)] = \phi(b)Pr(X=b)$.

Specializing now to the discrete case, we may say that the most general such case can be described as follows: There is a finite or enumerable number of points $(x_{1j},x_{2j},\ldots,x_{kj})$, $j = 1,2,\ldots$, and associated positive numbers p_j, $\sum_j p_j = 1$, such that

$$F(x_1,\ldots,x_k) = \sum p_i \text{ summed over all } i \text{ such that } x_{i1} \leq x_1,\ldots,x_{ik} \leq x_k.$$

In this case

$$\int_E \phi dF = \sum \phi(x_{1s},\ldots,x_{ks})\, p_s \text{ summed over all } s \text{ such that } (x_{1s},\ldots,x_{ks}) \in E.$$

In the continuous case

$$\int_E \phi dF = \int_E \phi f dV,$$

where dV is the volume element $dx_1dx_2\ldots dx_k$. In the most general case

$$\int_E dF = Pr(X \in E).$$

It is helpful for some of us to develop an intuitive feeling for the Stieltjes integral. Consider first an ordinary integral

$$\int_E h(x_1,\ldots,x_k)dV,$$

where h is continuous. We may conceive of the integral in a Leibnitzian (non-rigorous, but sometimes fruitful) sense: The k-dimensional volume E is partitioned into tiny volume elements dV. These are so small that the function h is "practically constant" over any dV. We multiply this "practically constant" value of h by the volume dV and sum over E. Now a c. d. f. $F(x_1,\ldots,x_k)$ defines a probability distribution over R_k, of

which it is sometimes convenient to think as a mass distribution. We think of dF as being the amount of mass or probability in an infinitesimal volume element dV, whether it be concentrated at points, along curves or surfaces, or smeared out as a density. We weight the "practically constant" value of ϕ in dV with the amount dF of mass or probability, getting ϕdF, and we sum over E. The reader may see that the definition of $\int_J \phi dF$ over a half-open cell J is a rigorous polishing up of the process we have described: In place of dV we use the cell J_j, in place of dF we use $\Delta_j^k F$, the probability that a random point be in J_j, we multiply not by the "practically constant" value of ϕ in J_j, but by any value it assumes in J_j, and finally, instead of merely summing, we take the limit of the sum.

2.6 Transformation of Variables

Suppose $y = \psi(x)$ is a (B-meas.) function of x. Then if X is a random variable with c. d. f. F(x), $Y = \psi(X)$ is also a random variable with c. d. f. G(y) calculated as follows:

$$G(y) = Pr(Y \leq y) = Pr(\psi(X) \leq y) = \int_{E_y} dF(x),$$

where E_y is the totality of points on the x-axis for which $\psi(x) \leq y$.

More generally, suppose $(X_1, X_2, \ldots, X_k)$ is a random vector variable with c.d. $F(x_1, x_2, \ldots, x_k)$, and $y_1, y_2, \ldots, y_n$ are (B-meas.) functions of $x_1, x_2, \ldots, x_k$, $y_i = \psi_i(x_1, x_2, \ldots, x_k)$. Then $(Y_1, Y_2, \ldots, Y_n)$, where $Y_i = \psi_i(X_1, X_2, \ldots, X_k)$, is a random vector variable with c. d. f.

$$G(y_1, y_2, \ldots, y_n) = \int_{E_{y_1, y_2, \ldots, y_n}} dF(x_1, x_2, \ldots, x_k),$$

where $E_{y_1, y_2, \ldots, y_n}$ is the region in R_k defined by $\psi_i(x_1, x_2, \ldots, x_k) \leq y_i$, $i = 1, 2, \ldots, n$.

It may be shown that if X_1, X_2 are random (possibly vector) variables, and that if $Y_1 = \psi_1(X_1)$, $Y_2 = \psi_2(X_2)$ are (B-meas.) functions, then if X_1 and X_2 are statistically independent, so are the random variables Y_1 and Y_2.

Transformations of discrete variables offer no especial difficulties, so we consider in the following sections transformations in the continuous case.

The theorems obtained there are essentially corollaries to corresponding theorems on the transformation of integrals, single and multiple. Rigorous proofs of the theorems on integrals may be found in standard real variable texts. For the student in this course the insertion at this point of heuristic proofs which will strengthen his

intuitive grasp seems desirable, and accordingly we employ the infinitesimal arguments so useful in applied mathematics.

2.61 Univariate Case

Suppose X is a random variable with p. d. f. $f(x)$. Let $y = \phi(x)$ be a monotone transformation having unique inverse $x = \phi^{-1}(y)$, and such that $\phi'(x)$ exists. Now consider a new random variable $Y = \phi(X)$. The problem here is to determine $Pr(y<\phi(X)<y+dy)$. Now since $y = \phi(x)$ is monotone, it is clear that the values of x for which $y < \phi(x) < y+dy (dy>0)$ will lie on an interval $(x,x+dx)$ depending on y, where dx may be positive or negative depending on whether $\phi(x)$ is monotone increasing or decreasing. Since $x = \phi^{-1}(y)$ by the inverse of the transformation $y = \phi(x)$, then expressed in terms of y, the interval $(x,x+dx)$ becomes $(\phi^{-1}(y), \phi^{-1}(y+dy))$. Hence the value of $Pr(y<\phi(X)<y+dy)$ is given by determining the value of $Pr(x<X<x+dx) = Pr(\phi^{-1}(y)<X<\phi^{-1}(y+dy))$ if $dx > 0$ or $Pr(x+dx<X<x) = Pr(\phi^{-1}(y+dy)<X<\phi^{-1}(y))$ if dx is negative. In either case the probability is, except for differentials of order higher than dy,

$$f(x)|dx| = f(x)\left|\frac{dx}{dy}\right|dy$$

where x is to be expressed in terms of y. We may summarize as follows:

Theorem (A): Let X be a continuous random variable with probability element $f(x)dx$, and let $y = \phi(x)$ be a monotone transformation with inverse $x = \phi^{-1}(y)$ such that $\phi'(x)$ exists. Then except for differentials of order higher than dy

$$Pr(y<\phi(X)<y+dy) = g(y)dy$$

where $g(y) = f(x)\left|\frac{dx}{dy}\right|$ expressed in terms of y.

Example. Suppose

$$f(x)dx = e^{-x}dx \quad (x\geq 0)$$
$$= 0\,dx \quad x<0$$

and that it is desired to find $Pr(y<X^2<y+dy)$, i. e., the probability element of y, say $g(y)dy$. We have the transformation $y = x^2$, or $x = \sqrt{y}$ and hence

$$g(y)dy = e^{-x}\left|\frac{dx}{dy}\right|dy = e^{-\sqrt{y}}\ \frac{1}{2\sqrt{y}}\,dy.$$

2.62 Bivariate Case

Suppose

(a) $$y_1 = y_1(x_1,x_2), \quad y_2 = y_2(x_1,x_2)$$

are functions of x_1, x_2 with continuous first partial derivatives. Let $f(x_1,x_2)$ be the joint p. d. f. of X_1 and X_2. We shall assume further that the transformation (a) is one-to-one, that is, the relation between the x's and y's is such that corresponding to

each point in the x_1x_2 plane (or that part of it for which the probability function $f(x_1,x_2) \neq 0$) there is one and only one point in the y_1y_2 plane and each point in the y_1y_2 plane which has a corresponding point in the x_1x_2 plane has one and only one corresponding point in the x_1x_2 plane, the relation between any point in the x_1x_2 plane and its corresponding point in the y_1y_2 plane being given by (a). Let the inverse of the transformation (a) be

$$x_1 = x_1(y_1,y_2), \qquad x_2 = x_2(y_1,y_2). \tag{b}$$

Let the Jacobian of the transformation (b) be

$$\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} \tag{c}$$

If X_1, X_2 are random variables, then $Y_1 = y_1(X_1,X_2)$ and $Y_2 = y_2(X_1,X_2)$ will also be random variables. The problem here is to determine the p. d. f. of Y_1 and Y_2, say $g(y_1,y_2)$, from $f(x_1,x_2)$ the p. d. f. of X_1, X_2 and the transformation (a). In other words, the problem is to determine

$$\Pr(y_1<Y_1<y_1+dy_1, y_2<Y_2<y_2+dy_2) \tag{d}$$

to within terms of order dy_1dy_2.

Consider the infinitesimal region R in the x_1,x_2 plane bounded by the curves whose equations are

$$\begin{aligned} y_1 &= y_1(x_1,x_2), & y_2 &= y_2(x_1,x_2), \\ y_1 + dy_1 &= y_1(x_1,x_2), & y_2+dy_2 &= y_2(x_1,x_2), \end{aligned} \tag{e}$$

where $dy_1>0$, $dy_2>0$.

The situation is represented in Figure 2.

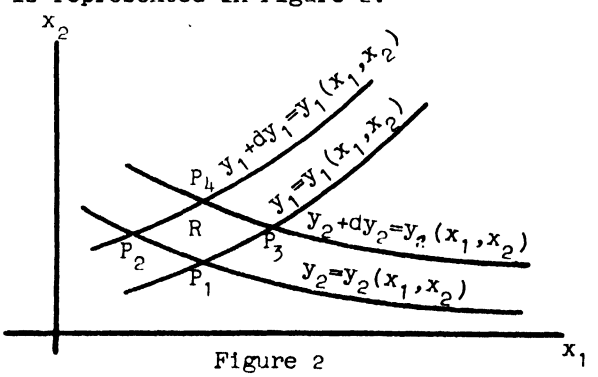

Figure 2

Now the probability (d) is given by $\iint_R f(x_1,x_2)dx_1dx_2$. By the mean value theorem for integrals the value of this integral is $f(x_1',x_2')dA$ where (x_1',x_2') is some point in R and dA is the area of R. We must now find an expression for dA.

If the coordinates of P_1 in Figure 2 are (x_1,x_2) then the coordinates of P_2, P_3, P_4 are

$$P_2: \quad \left(x_1+\frac{\partial x_1}{\partial y_1}\,dy_1,\ x_2+\frac{\partial x_2}{\partial y_1}\,dy_1\right), \quad P_3: \quad \left(x_1+\frac{\partial x_1}{\partial y_2}\,dy_2,\ x_2+\frac{\partial x_2}{\partial y_2}\,dy_2\right),$$

(f)

$$P_4: \quad \left(x_1+\frac{\partial x_1}{\partial y_1}\,dy_1+\frac{\partial x_1}{\partial y_2}\,dy_2,\ x_2+\frac{\partial x_2}{\partial y_1}\,dy_1+\frac{\partial x_2}{\partial y_2}\,dy_2\right)$$

except for infinitesimals of order higher than dy_1 and dy_2. To show this it is sufficient to consider only one point, say P_2. The coordinates of P_2 are given by (f) when y_1 is replaced by $y_1 + dy_1$. We have

$$x_1 = x_1(y_1+dy_1,y_2)$$

$$x_2 = x_2(y_1+dy_1,y_2).$$

But $x_1(y_1+dy_1,y_2) = x_1(y_1,y_2) + \frac{\partial x_1}{\partial y_1}dy_1$ + terms of order $(dy_1)^2$ and higher and $x_2(y_1+dy_1,y_2)$ $= x_2(y_1,y_2) + \frac{\partial x_2}{\partial y_1}\,dy_1$ + terms of order $(dy_1)^2$ and higher. But $(x_1(y_1,y_2),\ x_2(y_1,y_2))$ are the coordinates of P_1 which have been indicated by (x_1,x_2), thus showing that the approximate coordinates of P_2 are those stated in (f). A similar argument holds for the approximate coordinates of P_3 and P_4.

It is clear that P_1, together with the points represented by the approximate coordinates of P_2, P_3, P_4 given by (f) form a parallelogram R'. Now it is known from coordinate geometry that if (x_1,x_2), (x_1',x_2'), (x_1'',x_2'') are three vertices of a parallelogram, then the area of the parallelogram is given by the absolute value of the determinant

$$\begin{vmatrix} 1 & x_1 & x_2 \\ 1 & x_1' & x_2' \\ 1 & x_1'' & x_2'' \end{vmatrix}.$$

Hence the area of the parallelogram R' is given by the absolute value of

*

$$(g)\qquad \begin{vmatrix} 1 & x_1 & x_2 \\ 1 & x_1+\dfrac{\partial x_1}{\partial y_1}\,dy_1 & x_2+\dfrac{\partial x_2}{\partial y_1}\,dy_1 \\ 1 & x_1+\dfrac{\partial x_1}{\partial y_2}\,dy_2 & x_2+\dfrac{\partial x_2}{\partial y_2}\,dy_2 \end{vmatrix} = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_2}{\partial y_1} \\ \dfrac{\partial x_1}{\partial y_2} & \dfrac{\partial x_2}{\partial y_2} \end{vmatrix} dy_1dy_2 = \frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\,dy_1dy_2.$$

But since the coordinates of the vertices of parallelogram R' differ from the corresponding coordinates of the corresponding vertices of R by terms of order higher than dy_1 or dy_2, it follows that the area of R (i. e., dA) differs from the area of R' by terms of order higher than $dy_1\,dy_2$.

Since $f(x_1,x_2)$, the p. d. f. of X_1, X_2, is continuous, we have that $f(x_1',x_2')$ differs from $f(x_1,x_2)$ by terms of order dy_1dy_2, where (x_1',x_2') is any point in R. Therefore we have the result that the probability expressed by (d) is equal to

$$(h)\qquad f(x_1,x_2)\left|\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\right| dy_1dy_2,$$

where the x's are to be expressed in terms of y's by (b). It may be verified by the reader that

$$\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)} = \left[\frac{\partial(y_1,y_2)}{\partial(x_1,x_2)}\right]^{-1}.$$

We may summarize in the following:

Theorem (B): Let X_1, X_2 be two continuous random variables with p. d. f. $f(x_1,x_2)$. Let $y_1 = y_1(x_1,x_2)$, $y_2 = y_2(x_1,x_2)$ be a transformation with a unique inverse $x_1 = x_1(y_1,y_2)$, $x_2 = x_2(y_1,y_2)$, such that the first partial derivatives of the y's with respect to the x's exist. If the random variables $y_1(X_1,X_2)$ and $y_2(X_1,X_2)$ are denoted by Y_1 and Y_2 respectively, then

$$\Pr(y_1<Y_1<y_1+dy_1;\ y_2<Y_2<y_2+dy_2) = f(x_1,x_2)\left|\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\right| dy_1dy_2,$$

where x_1 and x_2 are expressed in terms of y_1,y_2 by (b), and $\dfrac{\partial(x_1,x_2)}{\partial(y_1,y_2)}$ is given by (c).

Example: To illustrate the transformation problem for two random variables, suppose the probability element of X_1 and X_2 is

$$f(x_1,x_2)dx_1dx_2 = \frac{1}{2\pi}\,e^{-\frac{1}{2}x_1^2-\frac{1}{2}x_2^2}\,dx_1dx_2$$

defined over the entire x_1, x_2 plane. To determine the p. d. f. of Y_1 and Y_2 where

$$Y_1 = \sqrt{X_1^2+X_2^2}, \quad Y_2 = \tan^{-1}\frac{X_1}{X_2}.$$

The transformation involved here is

$$y_1 = \sqrt{x_1^2+x_2^2}$$
$$y_2 = \tan^{-1}\frac{x_1}{x_2}$$

defined over that part of the y_1, y_2 plane for which $y_1 \geq 0$, $0 \leq y_2 < 2\pi$. The inverse of the transformation is

$$x_1 = y_1 \cos y_2$$
$$x_2 = y_1 \sin y_2.$$

We have

$$\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)} = \begin{vmatrix} \cos y_2 & -y_1 \sin y_2 \\ \sin y_2 & y_1 \cos y_2 \end{vmatrix} = y_1.$$

Therefore by Theorem (B), the probability element of Y_1, Y_2 is

$$\frac{1}{2\pi} e^{-\frac{1}{2}y_1^2} y_1 dy_1 dy_2.$$

2.63 k-Variate Case

Let the joint p. d. f. of $X_1, X_2, \ldots, X_k$ be $f(x_1,x_2,\ldots,x_k)$, and introduce new random variables $Y_1, Y_2, \ldots, Y_k$ by means of the one-to-one transformation

(a) $$y_i = y_i(x_1,x_2,\ldots,x_k), \qquad i = 1,2,\ldots,k.$$

Let the inverse (which will be unique) of this transformation be

(b) $$x_i = x_i(y_1,y_2,\ldots,y_k), \qquad i = 1,2,\ldots,k,$$

and its Jacobian

(c) $$\frac{\partial(x_1,x_2,\ldots,x_k)}{\partial(y_1,y_2,\ldots,y_k)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_k} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial x_k}{\partial y_1} & \frac{\partial x_k}{\partial y_2} & \cdots & \frac{\partial x_k}{\partial y_k} \end{vmatrix},$$

assuming, of course, that the first partial derivatives exist.

By pursuing an argument similar to that used in the bivariate case, we find that the probability element of the Y_1, say $g(y_1, y_2,\ldots,y_k)\, dy_1\ldots dy_k$, is given by

$$\text{(d)} \qquad f(x_1,x_2,\ldots,x_k) \left| \frac{\partial(x_1,x_2,\ldots,x_k)}{\partial(y_1,y_2,\ldots,y_k)} \right| dy_1 dy_2\ldots dy_k,$$

where the x's are to be expressed in terms of y's by (b).

This covers the cases where the number n of new variables equals the number k of original variables. It can be shown that if $n > k$, there exists no p. d. f. for the n new variables. (Note here the complete generality of the treatment by means of the c. d. f. in §2.6). If $n < k$ the usual method of getting the p. d. f. of the new variables is to adjoin further variables to fill out the number of new variables to k, use the above procedure, and then "integrate out" the extra variables by getting the marginal distribution of the n variables whose p. d. f. is desired.

2.7 Mean Value

We begin with the definition of the mean value of a random variable in general and then consider in later sections the mean values of particular (random) functions of especial interest in statistics. If X is a random variable with c. d. f. F(x) we define the mean value of X as

$$\text{(a)} \qquad E(X) = \int_{-\infty}^{+\infty} x\,dF(x).$$

This is also called the expected value of X.

If $Y = \phi(X)$ is a continuous function of X, then the c. d. f. of Y is (§2.6)

$$G(y) = \int_{E_y} dF(x),$$

where E_y is the set of points on the x-axis such that $\phi(x) \leq y$. From (a),

$$E(Y) = \int_{-\infty}^{+\infty} y\, dG(y),$$

and this may be shown to be equivalent to

$$\text{(b)} \qquad E[\phi(X)] = \int_{-\infty}^{+\infty} \phi(x)\,dF(x).$$

If random variables $X_1, X_2,\ldots,X_k$ have the c. d. f. $F(x_1,x_2,\ldots,x_k)$, and $y = \phi(x_1,x_2,\ldots,x_k)$ is continuous, then from the definition (a) it may be shown that

(c) $$E[\phi(X_1,X_2,\ldots,X_k)] = \int_{R_k}\phi dF,$$

where R_k is the entire k-space. Of course if the improper integral does not exist in the sense explained in §2.5, ff., we say that the mean value of ϕ does not exist. In the light of the intuitive discussion (§2.53) of the meaning of a Stieltjes integral, we see from (c) that the mean value of ϕ may be regarded as an average over k-space of the function ϕ,--the average being taken over volume elements dV, and the weight assigned to each contribution being the total probability in dV.

For the discrete and continuous cases, the expressions (b) and (c) may be analyzed into the forms given in §§2.51, 2.53.

2.71 Univariate Case; Tchebycheff's Inequality

The mean value of X^i,

$$\mu_i' = E(X^i) = \int_{-\infty}^{+\infty} x^i dF(x), \qquad i = 0,1,2,\ldots,$$

is called the i-th moment of the distribution F(x) about the origin. $\mu_0' = 1$ for any F(x); $\mu_1' = E(X)$ is called the mean of X, also the mean of the distribution, and denoted by a. The i-th moment about the mean is defined to be

(a) $$\mu_i = E[(X-a)^i] = \int_{-\infty}^{+\infty}(x-a)^i dF(x), \qquad i = 0,1,2,\ldots$$

For any F(x), $\mu_0 = 1$, $\mu_1 = 0$. The variance of X, or the variance of the distribution, is defined to be μ_2, and is denoted by the special symbol σ_X^2. $\sigma_X > 0$ is called the standard deviation of X or of the distribution. A formula for expressing μ_i in terms of μ_i', $\mu_{i-1}',\ldots,\mu_1'$ may be obtained by using the binomial theorem in (a) and then integrating In particular, we find that

$$\sigma_X^2 = \mu_2' - a^2.$$

An important theorem about arbitrary distributions with finite variance is contained in the Tchebycheff inequality:

(b) $$Pr(|X-a|>\delta\sigma_X) \leqq 1/\delta^2.$$

To prove (b) we break up the integral for

(c) $$\sigma_x^2 = \int_{-\infty}^{+\infty} (x-a)^2 dF(x) = \int_{I_1} + \int_{I_2} + \int_{I_3},$$

where the intervals I_1, I_2, I_3 are defined by

$$I_1: \ -\infty < x < a - \delta\sigma_x,$$
$$I_2: \ a - \delta\sigma_x \leqq x \leqq a + \delta\sigma_x,$$
$$I_3: \ a + \delta\sigma_x < x < +\infty.$$

Now in I_1

$$|x-a| > \delta\sigma,$$

Hence

(d) $$\int_{I_1} (x-a)^2 dF(x) \geqq \delta^2\sigma_x^2 \int_{I_1} dF(x).$$

Similarly,

(e) $$\int_{I_3} (x-a)^2 dF(x) \geqq \delta^2\sigma_x^2 \int_{I_3} dF(x).$$

Finally,

(f) $$\int_{I_2} (x-a)^2 dF(x) \geqq 0.$$

Using (d), (e), (f) in (c), we get

$$\delta^2\sigma_x^2 \int_{I_1+I_3} dF(x) \leqq \sigma_x^2.$$

This is easily seen to be equivalent to (b).

2.72 Bivariate Case

For the distribution $F(x_1,x_2)$ we define moments μ'_{ij} about the origin by

$$\mu'_{ij} = E(X_1^i X_2^j) = \int_{R_2} x_1^i x_2^j dF(x_1,x_2), \qquad i,j = 0,1,2..,$$

where R_2 is the entire x_1x_2-space. Since X_1 has the marginal distribution $F_1(x_1)$, the mean of X_1 has already been defined in §2.71; we denote it by a_1. In view of the remarks in §2.7, we may calculate a_1 from either of the integrals

$$a_1 = \int_{-\infty}^{+\infty} x_1 dF_1(x_1) = \int_{R_2} x_1 dF(x_1,x_2) = \mu'_{10}.$$

Similar statements apply to $a_2 = E(x_2)$. We note $\mu'_{00} = 1$. The point (a_1,a_2) may be called the mean of the distribution, the moments μ_{ij} about the mean for $F(x_1,x_2)$ are defined by

(a) $$\mu_{ij} = E[(X_1-a_1)^i(X_2-a_2)^j] = \int_{R_2}(x_1-a_1)^i(x_2-a_2)^j dF(x_1,x_2), \qquad i,j = 0,1,2,\ldots$$

For any $F(x_1,x_2)$, $\mu_{00} = 1$, $\mu_{10} = \mu_{01} = 0$. The variance of X_1 has already been defined in §2.71; we note that it is $\sigma^2_{x_1} = \mu_{20}$. Likewise, $\sigma^2_{x_2} = \mu_{02}$. The remaining second order moment μ_{11} is called the covariance of X_1 and X_2. The quotient

(b) $$\rho_{12} = \mu_{11} / \sigma_{x_1}\sigma_{x_2}$$

is called the correlation coefficient of X_1 and X_2. By means of the Schwartz inequality it may be shown that $-1 \leq \rho_{12} \leq 1$. As an exercise the reader may show that if X_1 and X_2 are statistically independent, then $\rho_{12} = 0$, but the converse is false.

The reader may also verify that a necessary and sufficient condition for $\rho_{12} = 1$ is that all of the probability in the X_1X_2 plane be concentrated along some straight line with positive slope. (For $\rho_{12} = -1$ the slope must be negative.)

Formulas giving the moments about the mean in terms of the moments about the origin may again be obtained from (a); in particular, it is found that

$$\mu_{11} = \mu'_{11} - a_1a_2,$$
$$\sigma^2_{x_1} = \mu'_{20} - a_1^2,$$
$$\sigma^2_{x_2} = \mu'_{02} - a_2^2,$$

and these expressions may then be substituted in (b) to evaluate the correlation coefficient in terms of the first and second order moments about the origin.

2.73 k-Variate Case

The moments μ' of a distribution $F(x_1,x_2,\ldots,x_k)$ about the origin are defined as

$$\mu'_{j_1j_2\cdots j_k} = E(X_1^{j_1}X_2^{j_2}\ldots X_k^{j_k}) = \int_{R_k}\prod_{i=1}^{k} x_1^{j_1} dF, \qquad j_1 = 0,1,2,\ldots,$$

where R_k is the complete k-space. For any F, $\mu'_{00\ldots0} = 1$. The mean of X_1, defined in §2.71, may now be seen to be $\mu'_{100\ldots0}$, and can be expressed also by means of integrals with respect to marginal distributions of various orders. We denote $E(X_1)$ by a_1, and note that the above statements apply to $a_2 = E(X_2),\ldots,a_k = E(X_k)$. The point $(a_1,a_2,\ldots,a_k)$ is called the mean of the distribution, and the moments μ about the mean are defined to be

$$\mu_{j_1 j_2 \cdots j_k} = E[\prod_{i=1}^{k}(X_i-a_i)^{j_i}] = \int_{R_k} \prod_{i=1}^{k}(x_i-a_i)^{j_i} dF, \qquad j_i = 0,1,2,\ldots$$

We note $\mu_{00\ldots0} = 1$. In order to simplify the notation, we specialize the following remarks to the variable X_1 or the pair X_1,X_2; their generalizations are obvious: $\mu_{100\ldots0} = 0$. The variance of X_1, defined in §2.71 is seen to be $\mu_{200\ldots0}$. The covariance of X_1 and X_2, defined in §2.72, is $\mu_{1100\ldots0}$, and the correlation coefficient of $X_1$and X_2 is

$$\rho_{12} = \mu_{1100\ldots0}/(\mu_{200\ldots0}\,\mu_{020\ldots0})^{\frac{1}{2}}.$$

These quantities may all be expressed in terms of the first and second order moments about the origin.

2.74 Mean and Variance of a Linear Combination of Random Variables

Suppose we have k random variables $X_1,X_2,\ldots,X_k$, the c. d. f. of X_1 being $F_1(x_1)$. Let their joint c. d. f. be $F(x_1,x_2,\ldots,x_k)$. $F_1(x_1)$ is then the marginal distribution (§2.2) of X_1; if the X_1 are mutually (statistically) independent

$$F(x_1,x_2,\ldots,x_k) = \prod_{i=1}^{k} F_1(x_1),$$

but we shall not assume this. Let $y = \psi(x_1,x_2,\ldots,x_k)$ be a linear function,

(a) $$y = \sum_{i=1}^{k}\alpha_i x_i.$$

Then $Y = \psi(X_1,X_2,\ldots,X_k) = \sum_{i=1}^{k}\alpha_i X_i$ is a random variable (§2.6), its c. d. f. G(y) is

$$G(y) = \int_{E_y} dF(x_1,x_2,\ldots,x_k),$$

where E_y is the half-space defined by $\sum_{i=1}^{k}\alpha_i x_i \leq y$.

In accordance with the notation established in §2.73, denote the mean of X_i by a_i, its variance by $\sigma^2_{X_i}$ which we shall now abbreviate to σ_i^2, and the covariance of X_i and X_j by $\rho_{ij}\sigma_i\sigma_j$. Denote the mean of Y by a, its variance by σ_y^2.

It is helpful to note that E is a linear operator: if ϕ_1 and ϕ_2 are continuous functions of $X_1, X_2, \ldots, X_k$, and A and B are constants,

$$E(A\phi_1+B\phi_2) = \int_{R_k}(A\phi_1+B\phi_2)dF = A\int_{R_k}\phi_1 dF + B\int_{R_k}\phi_2 dF = A\,E(\phi_1) + B\,E(\phi_2).$$

From this we get immediately, because of (a),

$$a = E(Y) = \sum_{i=1}^{k}\alpha_i E(X_i) = \sum_{i=1}^{k}\alpha_i a_i. \tag{b}$$

Note that for the validity of this result it is irrevelant whether or not the X_i are statistically independent.

Next let us calculate the variance of Y:

$$\sigma_y^2 = E\{(Y-a)^2\} = E\{[\sum_{i=1}^{k}\alpha_i X_i - \sum_{i=1}^{k}\alpha_i a_i]^2\}$$

$$= E\{[\sum_{i=1}^{k}\alpha_i(X_i-a_i)]^2\} = E\{\sum_{i,j=1}^{k}\alpha_i\alpha_j(X_i-a_i)(X_j-a_j)\}$$

$$= \sum_{i,j=1}^{k}\alpha_i\alpha_j E[(X_i-a_i)(X_j-a_j)]$$

$$\sigma_y^2 = \sum_{i,j=1}^{k}\alpha_i\alpha_j\sigma_i\sigma_j\rho_{ij}, \tag{c}$$

where $\rho_{ii} = 1$. If the X_i are mutually independent, then $\rho_{ij} = 0$ for $i \neq j$, and

$$\sigma_y^2 = \sum_{i=1}^{k}\alpha_i^2\sigma_i^2. \tag{d}$$

2.75 Covariance and Correlation between two Linear Combinations of Random Variables

Suppose Y_α and Y_β are each linear combinations of random variables. The random variables in both combinations may be the same, or none of those appearing in Y_α may appear in Y_β, or there may be an intermediate degree of overlapping. All of these cases may be covered by assuming that

$$Y_\alpha = \sum_{i=1}^{k} \alpha_i X_i, \qquad Y_\beta = \sum_{i=1}^{k} \beta_i X_i,$$

where the α_i, β_i are constants and the X_i are random variables with joint c. d. f. $F(x_1,x_2,\ldots,x_k)$. For example, the case of no overlapping would be obtained by requiring $\alpha_i\beta_i = 0$, $i = 1,2,\ldots,k$. If $E(X_i) = a_i$, then from (b) of §2.74,

$$E(Y_\alpha) = \sum_{i=1}^{k} \alpha_i a_i, \qquad E(Y_\beta) = \sum_{i=1}^{k} \beta_i a_i.$$

Hence the covariance of Y_α and Y_β is

$$E\{[Y_\alpha - E(Y_\alpha)][Y_\beta - E(Y_\beta)]\} = E\{[\sum_{i=1}^{k} \alpha_i (X_i - a_i)][\sum_{j=1}^{k} \beta_j (X_j - a_j)]\} = \sum_{i,j=1}^{k} \alpha_i \beta_j \sigma_i \sigma_j \rho_{ij},$$

where σ_i^2 is the variance of X_i and $\sigma_i\sigma_j\rho_{ij}$ is the covariance X_i and X_j. Hence the correlation coefficient between Y_α and Y_β is

(a)
$$\rho_{Y_\alpha Y_\beta} = \frac{\sum_{i,j=1}^{k} \alpha_i \beta_j \sigma_i \sigma_j \rho_{ij}}{\sqrt{\sum_{i,j=1}^{k} \alpha_i \alpha_j \sigma_i \sigma_j \rho_{ij}} \sqrt{\sum_{i,j=1}^{k} \beta_i \beta_j \sigma_i \sigma_j \rho_{ij}}},$$

from (b) of §2.72 and (c) of §2.74. Special cases of this formula for the correlation coefficient are much used in education and psychology in connection with tests.

2.76 The Moment Problem

The general moment problem (univariate) is twofold: (i) given an infinite sequence of numbers 1, μ_1', μ_2',..., does there exist a distribution with these numbers as moments? and if so, (ii) is the distribution unique? It is usually only the problem (ii) that arises in statistics. It may be shown that whenever the moment generating function $\phi(\theta)$ (see §2.8) exists for $-h \leq \theta \leq h$, $h > 0$, there is a unique* distribution with the moments $\phi^{(i)}(0)$.

Necessary and sufficient conditions for the unique determination of a

*ϕ then is analytic in a strip containing the imaginary axis, hence the characteristic function $\bar{\phi}(t) = \phi(it)$ is analytic for all real t, and this is a sufficient condition for uniqueness in the moment problem: See P. Lévy, Theorie de l'addition des variables aleatoires, Monographies des probabilités, Paris, 1937, p. 41.

distribution by its moments are extremely complicated,* but the following theorem** gives an easily applied sufficient condition of Carleman:

Theorem (A): A sufficient condition for the uniqueness of a distribution with moments μ_i' is that the series $\sum_{m=1}^{\infty}(\mu_{2m}')^{-1/2m}$ diverge.

For a multivariate distribution with moments μ' define

(a) $$\lambda_1 = \mu'_{100\ldots0} + \mu'_{0100\ldots0} + \cdots + \mu'_{0\ldots001}.$$

A sufficient condition of Cramér and Wold** for uniqueness is Theorem (B), of which (A) may be regarded as a special case:

Theorem (B): If the series $\sum_{m=1}^{\infty}(\lambda_{2m})^{-1/2m}$ diverges, where λ_1 is defined by (a), then the distribution $F(x_1,x_2,\ldots,x_k)$ is uniquely determined by its moments.

2.8 Moment Generating Functions

When the moment generating function (m. g. f.) of a distribution satisfies a certain condition given below, then the moments of the distribution may easily be found by differentiation of the moment generating function. The use of the m. g. f. also permits the easy determination of the distribution of certain functions of certain random variables. We consider in detail the

2.81 Univariate Case

For any distribution $F(x)$ we define the m. g. f. as

(a) $$\phi(\theta) = E(e^{\theta X}) = \int_{-\infty}^{+\infty} e^{\theta x}dF(x).$$

If we proceed heuristically, we may write

(b) $$\phi^{(1)}(0) = \left[\frac{d^1}{d\theta^1}\int_{-\infty}^{+\infty} e^{\theta x}dF(x)\right]_{\theta=0} = \int_{-\infty}^{+\infty}\left[\frac{d^1}{d\theta^1}e^{\theta x}\right]_{\theta=0}dF(x) = \int_{-\infty}^{+\infty} x^1 dF(x) = \mu_1'.$$

Let us now consider under what conditions $\mu_1' = \phi^{(1)}(0)$.

In order that $\phi(\theta)$, considered as a function of a real variable, possess derivatives at $\theta = 0$, it is necessary that $\phi(\theta)$ as defined by (a) exist in a neighborhood

---

*H. Hamburger, "Über eine Erweiterung des Stieltjesschen Momentenproblems", Math. Annalen, vol. 81 (1920), pp. 235-319, and vol. 82 (1921), pp. 120-164, 168-187.

**H. Cramér and H. Wold, "Some theorems on distribution functions", Jour. London Math. Soc., vol. 11 (1936), pp. 290-294.

$-h \leq \theta \leq h$, $h > 0$. (Note that in any case $\phi(0) = 1$ is defined by (a)). We see now that this restricts the class of functions $F(x)$ under consideration. Our definition (§2.51) of the infinite integral $\int_{-\infty}^{+\infty}$ implies the existence of $\int_{0}^{+\infty}$ and $\int_{-\infty}^{0}$. Hence as $x \to +\infty$, $F(x) \to 1$ sufficiently rapidly so that

$$(c) \qquad M_1 = \int_{0}^{+\infty} e^{hx} dF(x) < \infty,$$

and as $x \to -\infty$, $F(x) \to 0$ sufficiently rapidly so that

$$(d) \qquad M_2 = \int_{-\infty}^{0} e^{-hx} dF(x) < \infty.$$

This means that $F(x)$ possesses moments of all orders: To demonstrate the finiteness of

$$\mu_i' = \int_{-\infty}^{+\infty} x^i dF(x),$$

consider

$$\int_{0}^{+\infty} x^i dF(x) = \int_{0}^{a} x^i dF(x) + \int_{a}^{+\infty} (x^i e^{-hx}) e^{hx} dF(x).$$

Choose a so large that $x^i e^{-hx} < 1$ for $x > a$. Then the second term of the right member is less than M_1 defined by (c); the first term is certainly finite, and thus

$$\int_{0}^{+\infty} x^i dF(x) < \infty.$$

Similarly by use of (d) we may show

$$\left| \int_{-\infty}^{0} x^i dF(x) \right| < \infty,$$

and hence $|\mu_i'| < \infty$ for all i.

We now state the heuristically obtained relation (b) in the form of

Theorem (A): If the m. g. f. $\phi(\theta)$ of a c. d. f. $F(x)$, as defined by (a), exists for $-h \leq \theta \leq h$, where $h > 0$, then the i-th moment of $F(x)$ about the origin is

$$\mu_i' = \phi^{(i)}(0), \qquad i = 0, 1, 2, \ldots$$

The proof of this theorem may be based on the theory of the bilateral Laplace transform and is beyond the level of this course.*

The m. g. f. if it exists is uniquely determined by (a). The converse** is stated in

Theorem (B): If $F(x)$ has the m. g. f. $\phi(\theta)$, and $\phi(\theta)$ exists for $-h \leq \theta \leq h$, $h > 0$, and if the c. d. f. $G(x)$ has the same m. g. f., then $G(x) \equiv F(x)$.

The reader may write out an expression for $\phi(\theta)$ in the discrete case, which is a sum of terms, and an expression in the continuous case, which is an ordinary integral, by using the analysis of §2.51.

We note that if $Y = \psi(X)$ is a continuous function of X, and $G(y)$ is the c. d. f. of Y, then the m. g. f. of $G(y)$ is

$$E(e^{\theta Y}) = E(e^{\theta\psi(X)}) = \int_{-\infty}^{+\infty} e^{\theta\psi(x)} dF(x).$$

If this exists for $|\theta| \leq h$ $(h>0)$ and is recognized as the m. g. f. of a known distribution, then theorem (B) determines $G(y)$.

In certain problems, particularly in sampling theory, it is important to know the limiting form of a c. d. f. $F_{(n)}(x)$ as $n \to \infty$ of a function X_n of n random variables. The m. g. f. offers a powerful method for determining the limit of this distribution. The method is to obtain the m. g. f. of X_n, say $\phi_n(\theta)$; then if $\phi_n(\theta)$ has a limiting form as $n \to \infty$ which is the m. g. f. of some c. d. f. $F(x)$, we may conclude under certain conditions that $\lim_{n \to \infty} F_{(n)}(x) = F(x)$. More precisely we shall state the following theorem without proof.***

Theorem (C): Let $F_{(n)}$ and $\phi_{(n)}(\theta)$ be respectively the c. d. f. and m. g. f. of a random variable $X_n (n=1,2,3,4,\ldots)$. If $\phi_{(n)}(\theta)$ exists for $|\theta| < h$ for all n and if there exists a function $\phi(\theta)$ such that $\lim_{n\to\infty} \phi_{(n)}(\theta) = \phi(\theta)$ for $|\theta| < h'$, then $\lim_{n\to\infty} F_{(n)}(x) = F(x)$, where $F(x)$ is the c. d. f. of a random variable X with m. g. f. $\phi(\theta)$.

---

*D. V. Widder, The Laplace Transform, Princeton University Press, 1941, p. 240.

**If the integral defining $\phi(\theta)$ exists on the real interval $(-h,h)$, it exists for complex θ in the strip determined by the condition that the real part of θ be in the interval, and ϕ is an analytic function in the strip: see Widder, loc. cit. Hence if for $F(x)$ and $G(x)$ the moment generating functions coincide in the interval, they coincide in the strip. For coincidence in the strip there is a uniqueness theorem: Widder, p. 243.

***For proof, see J. H. Curtiss, "On the Theory of Moment Generating Functions", Annals of Math. Stat., Vol. 13, No. 4, pp. 430-433.

2.82 Multivariate Case

The m. g. f. of a distribution $F(x_1,x_2,\ldots,x_k)$ is defined to be

$$\text{(a)} \qquad \phi(\theta_1,\theta_2,\ldots,\theta_k) = E(e^{\sum_{i=1}^{k}\theta_i X_i}) = \int_{R_k} e^{\sum_{i=1}^{k}\theta_i x_i}\, dF.$$

We assume

$$\text{(b)} \qquad \phi \text{ exists for } -h \leqq \theta_i \leqq h,\ h > 0, \qquad i = 1,2,\ldots,k,$$

and then may consider restrictions on $F(x)$, analogous to those of §2.81, implied by (b). We state without proof

Theorem (A): Under the assumption (b)

$$\mu'_{j_1 j_2 \ldots j_k} = \left[\frac{\partial^{j_1+j_2+\ldots+j_k}}{\partial\theta_1^{j_1}\partial\theta_2^{j_2}\ldots\partial\theta_k^{j_k}}\,\phi(\theta_1,\theta_2,\ldots,\theta_k)\right]_{\theta_1=\theta_2=\ldots=\theta_k=0}.$$

Theorem (B): If ϕ satisfies condition (b), it uniquely determines F.

Let $F_i(x_i)$, with m. g. f. $\phi_i(\theta_i)$, be the c. d. f.'s of mutually independent variables X_i, $i = 1,2,\ldots,k$. Then the joint c. d. f. is

$$\text{(c)} \qquad F(x_1,x_2,\ldots,x_k) = \prod_{i=1}^{k} F_i(x_i),$$

and the m. g. f. of F is

$$\phi(\theta_1,\theta_2,\ldots,\theta_k) = \int_{R_k} e^{\sum_{i=1}^{k}\theta_i x_i}\, dF = \prod_{i=1}^{k}\int_{-\infty}^{+\infty} e^{\theta_i x_i}\, dF_i(x_i),$$

$$\text{(d)} \qquad \phi(\theta_1,\theta_2,\ldots,\theta_k) = \prod_{i=1}^{k}\phi_i(\theta_i).$$

By the uniqueness Theorem (B) it follows that if the m. g. f. is (d), the distribution is (c).

Theorem (C): Suppose that random variables X_i, $i = 1,2,\ldots,k$, have c. d. f.'s $F_i(x_i)$ with m. g. f. $\phi_i(\theta_i)$, and that all $\phi_i(\theta_i)$ satisfy condition (b). Then the X_i are mutually independent if and only if the m. g. f. ϕ of the joint distribution F factors according to (d).

The theorem is also valid in the case where the X_i are vector variables (then θ_i are also vectors).

If $Y_1 = \psi_1(X_1,X_2,\ldots,X_k)$, $i = 1,2,\ldots,t$, are continuous functions, then a method of determining the joint c. d. f. $G(y_1,y_2,\ldots,y_t)$ of the variables Y_i is to form the m. g. f. of G; it is

$$\phi(\theta_1,\theta_2,\ldots,\theta_t) = E\left[e^{\sum_{i=1}^{t}\theta_i Y_i}\right] = E\left[e^{\sum_{i=1}^{t}\theta_i\psi_i(X_1,X_2,\ldots,X_k)}\right]$$

$$= \int_{R_k} e^{\sum_{i=1}^{t}\theta_i\psi_i(x_1,x_2,\ldots,x_k)}\, dF(x_1,x_2,\ldots,x_k).$$

If this exists for $|\theta_i| \leq h > 0$, $i = 1,2,\ldots,t$, it uniquely determines $G(y_1,y_2,\ldots,y_t)$.

2.9 Regression

2.91 Regression Functions

If X_1, X_2 have the joint p. d. f. $f(x_1,x_2)$, we define the regression function $a_{1\cdot x_2}$ of X_1 on X_2 as the mean value of X_1 for a fixed value x_2 of X_2, i. e.

$$\text{(a)} \qquad a_{1\cdot x_2} = E(X_1|X_2=x_2) = \int_{-\infty}^{+\infty} x_1 f(x_1|x_2)\,dx_1,$$

where the conditional p. d. f. $f(x_1|x_2)$ is defined by (g) of §2.4. We note that the regression function (a) is a function of x_2 only. The graph of this function is called the regression curve. If the regression function is linear,

$$\text{(b)} \qquad a_{1\cdot x_2} = E(X_1|X_2=x_2) = bx_2 + c,$$

then we say that we have a case of linear regression, and call b and c the regression coefficients. The reader may show that if X_1 and X_2 are statistically independent, then the regression of X_1 on X_2 is linear, with $b = 0$ and $c = a_1$, the mean of X_1. We remark that the regression of X_1 on X_2 may be linear, while that of X_2 on X_1 is not.

If X_1, X_2 are discrete random variables, then in the notation of §2.12, we define the regression of X_1 on X_2 only for $X_2 = x_{2i}$, $i = 1,2,\ldots$, by

$$\text{(c)} \qquad a_{1\cdot x_2} = E(X_1|X_2=x_{2i}) = \sum p_j x_{1j} / \sum p_j,$$

where both summations are made for all j such that $x_{2j} = x_{2i}$. For the mixed case described in §2.12, we define the regression of X_1 on X_2 by

(d) $$a_{1\cdot x_2} = E(X_1|X_2=x_{21}) = \int_{-\infty}^{+\infty} x_1 f(x_1|x_{21})dx_1.$$

We shall limit the discussion for more than two variables to the continuous case. For k random variables $X_1,X_2,\ldots,X_k$, let $f(x_1|x_2,x_3,\ldots,x_k)$ be the conditional p. d. f. defined by (h) of §2.4. Then we define the regression function of X_1 on $X_2,X_3,\ldots,X_k$ to be

(e) $$a_{1\cdot x_2x_3\ldots x_k} = E(X_1|X_i=x_i, i=2,3,\ldots,k) = \int_{-\infty}^{+\infty} x_1 f(x_1|x_2,x_3,\ldots,x_k)dx_1.$$

If this function of $x_2,x_3,\ldots,x_k$ is linear,

(f) $$a_{1\cdot x_2x_3\ldots x_k} = E(X_1|X_1=x_1) = \sum_{j=2}^{k} b_j x_j + c,$$

then the regression is said to be linear and the b_j and c are called regression coefficients. Similarly, we may define the regression function of any X on the remaining X's. We note in conclusion that a regression function may always be regarded as the first moment of a conditional distribution.

2.92 Variance about Regression Functions

The variance of X_1 for a fixed value x_2 of X_2 is defined as

(a) $$\sigma^2_{1\cdot x_2} = \int_{-\infty}^{\infty} (x_1-a_{1\cdot x_2})^2 f(x_1|x_2)dx_1.$$

$\sigma^2_{1\cdot x_2}$ is, in general, a function of x_2, and its mean value $\sigma^2_{1\cdot 2}$ with respect to x_2 is known as the variance of X_1 about the regression function of X_1 on X_2. That is, we have

(b) $$\sigma^2_{1\cdot 2} = \int_{-\infty}^{\infty} \sigma^2_{1\cdot x_2} f_2(x_2)dx_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1-a_{1\cdot x_2})^2 f(x_1,x_2)dx_1dx_2.$$

In the k-variate case, we have

(c) $$\sigma^2_{1\cdot x_2x_3\ldots x_k} = \int_{-\infty}^{\infty} (x_1-a_{1\cdot x_2x_3\ldots x_k})^2 f(x_1|x_2x_3\ldots x_k)dx_1,$$

and the variance of X_1 about the regression function of X_1 on $X_2,X_3,\ldots,X_k$ is

(d) $$\sigma^2_{1\cdot 23\ldots k} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \sigma^2_{1\cdot x_2x_3\ldots x_k} f(x_2,x_3,\ldots,x_k)dx_2\ldots dx_k$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(x_1-a_{1\cdot x_2x_3\ldots x_k})^2 f(x_1,x_2,\ldots,x_k)dx_1dx_2\ldots dx_k .$$

The quantities given by (a), (b), (c) and (d) may be similarly defined for discrete and mixed cases, and also for empirical distributions.

2.93 Partial Correlation

Suppose $X_1, X_2,\ldots,X_k$ is a set of random variables. The covariance between any two of the variables, say X_1 and X_2, for fixed values of any set of the remaining variables, say $X_r, X_{r+1},\ldots,X_k$ $(2<r\leq k)$, is defined as

(a) $$c_{12\cdot x_rx_{r+1}\ldots x_k} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(x_1-a_{1\cdot x_rx_{r+1}\ldots x_k})(x_2-a_{2\cdot x_rx_{r+1}\ldots x_k})f(x_1x_2|x_r\ldots x_k)dx_1dx_2.$$

Let $c_{12\cdot r(r+1)\ldots k}$ be the mean value of $c_{12\cdot x_rx_{r+1}\ldots x_k}$ with respect to $x_r,x_{r+1},\ldots,x_k$

(b) $$c_{12\cdot r(r+1)\ldots k} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} c_{12\cdot x_r\,x_{r+1}\ldots x_k} f(x_rx_{r+1}\ldots x_k)dx_r\ldots dx_k$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(x_1-a_{1\cdot x_rx_{r+1}\ldots x_k})(x_2-a_{2\cdot x_rx_{r+1}\ldots x_k})f(x_1,x_2,x_r,\ldots x_k)dx_1dx_2dx_r\ldots dx_k$$

The partial correlation coefficient $\rho_{12\cdot r(r+1)\ldots k}$ between X_1, X_2 with respect to X_r, $X_{r+1},\ldots,X_k$ is defined as

(c) $$\rho_{12\cdot r(r+1)\ldots k} = \frac{c_{12\cdot r(r+1)\ldots k}}{\sigma_{1\cdot r(r+1)\ldots k}\,\sigma_{2\cdot r(r+1)\ldots k}}$$

The quantities defined in (a), (b) and (c) extend to discrete and mixed cases.

2.94 Multiple Correlation

A procedure which is often carried out in statistics is that of determining best-fitting linear regression functions in the sense of least squares even though the actual regression function is "not quite" linear. The procedure is perhaps more often carried out with an empirical c. d. f. $F_n(x_1,x_2,\ldots,x_k)$ than with a probability c. d. f.

Here, we shall only consider the case of a probability c. d. f. where the variables are all continuous. There will be analagous results for discrete and mixed cases (and also for the empirical distributions).

In this problem we let $X_1, X_2,..,X_k$ be random variables with c. d. f. $F(x_1,x_2,..,x_k)$ and determine the constants $b_1, b_2,..,b_k$ so that the mean value of the square $(x_1-b_1-\sum_{i=2}^{k} b_i x_i)^2$ is a minimum, i. e., so that

$$(a) \qquad S = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} (x_1-b_1-\sum_{i=2}^{k} b_i x_i)^2 dF(x_1,x_2,..,x_k) = E[(X_1-b_1-\sum_{i=2}^{k} b_i X_i)^2]$$

is a minimum.

The values of the b's which minimize S are given by solving the equations $\frac{\partial S}{\partial b_i} = 0$, $(i=1,2,\ldots,k)$. Writing out these equations we have (after dividing each equation by -2):

$$(b) \qquad \begin{aligned} a_1-b_1-b_2a_2-b_3a_3-\ldots-b_ka_k &= 0,\\ c_{12}-b_1a_2-b_2c_{22}-b_3c_{23}-\ldots-b_kc_{2k} &= 0,\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots &\\ c_{1k}-b_1a_k-b_2c_{k2}-b_3c_{k3}-\ldots-b_kc_{kk} &= 0, \end{aligned}$$

where $a_i = E(X_i)$ and $c_{ij} = E(X_iX_j)$. Substituting the value of b_1 from the first equation into each of the remaining equations, and setting $C_{ij} = c_{ij}-a_ia_j = E[(X_i-a_i)(X_j-a_j)]$, the covariance between X_i and X_j, we have the following equations to solve for $b_2, b_3,..,b_k$:

$$(c) \qquad \sum_{i=2}^{k} b_i C_{ij} = C_{1j} \qquad (j = 2,3,\ldots,k)$$

from which we obtain by using Cramér's rule for solving linear equations

$$(d) \qquad b_i = \sum_{j=2}^{k} C_{1j}C^{ij}$$

where $C^{ij} = \dfrac{\text{cofactor of } C_{ij} \text{ in } |C_{ij}|}{|C_{ij}|}$,

$|C_{ij}|$ being the determinant $\begin{vmatrix} C_{22} & C_{23} & \cdots & C_{2k} \\ C_{32} & C_{33} & & C_{3k} \\ \vdots & & \ddots & \vdots \\ C_{k2} & C_{k3} & \cdots & C_{kk} \end{vmatrix}$

It is assumed, of course, that this determinant $\neq 0$.

For the value of b_1 we therefore have

$$(e) \qquad b_1 = a_1 - \sum_{i=2}^{k} b_i a_i = a_1 - \sum_{i,j=2}^{k} a_i C_{1j} C^{ij} .$$

The <u>least squares</u> regression function of X_1 on $X_2, X_3, \ldots, X_k$ is thus

$$(f) \qquad b_1 + \sum_{i=2}^{k} b_i X_i ,$$

where the values of the b's are given by (d) and (e).

If we substitute the minimizing values of the b's, given by (d) and (e), in (a) we obtain the minimum value of S:

$$(g) \qquad \begin{aligned} \text{Min}(S) = E[(X_1 - a_1 - \sum_{i,j=2}^{k} (X_i - a_i) C_{1j} C^{ij})^2] = C_{11} - 2 \sum_{i,j=2}^{k} C_{11} C_{1j} C^{ij} \\ + \sum_{i',j'=2}^{k} \sum_{i,j=2}^{k} C_{1j} C_{1j'} C_{ii'} C^{ij} C^{i'j'} . \end{aligned}$$

If we sum the last expression first with respect to i, we find that $\sum_{i=2}^{k} C_{ii'} C^{ij} = 1$ if $i' = j$, and $= 0$ if $i' \neq j$. Hence summing on i and putting $i' = j$ the last expression reduces to $\sum_{j,j'=2}^{k} C_{1j} C_{1j'} C^{jj'}$ which is the same as $\sum_{i,j=2}^{k} C_{11} C_{1j} C^{ij}$. Thus denoting Min(S) by $\bar{\sigma}^2_{1 \cdot 2\,3 \ldots k}$ we have

$$(h) \qquad \bar{\sigma}^2_{1 \cdot 23 \ldots k} = C_{11} - \sum_{i,j=2}^{k} C_{11} C_{1j} C^{ij}$$

$$= \frac{\begin{vmatrix} C_{11} & C_{12} & \cdots & C_{1k} \\ C_{21} & C_{22} & \cdots & C_{2k} \\ \vdots & \vdots & & \vdots \\ C_{k1} & C_{k2} & \cdots & C_{kk} \end{vmatrix}}{|C_{ij}|}$$

To show that $\bar{\sigma}^2_{1 \cdot 23 \ldots k}$ may be expressed as this ratio of determinants, let us note that the determinant in the numerator may be expressed as

$$(i) \qquad C_{11} \bar{C}_{11} + C_{12} \bar{C}_{12} + \cdots + C_{1k} \bar{C}_{1k} ,$$

where $\overline{C}_{11}$ = cofactor of C_{11} in the numerator determinant. Now, for i = 2,3,..,k,

(j) $$\overline{C}_{11} = -(C_{12}\overline{C}_{12}+C_{13}\overline{C}_{13}+\ldots+C_{1k}\overline{C}_{1k})$$

where $\overline{C}_{ij}$ is the cofactor of C_{ij} in the determinant $|C_{ij}|$, (i,j=2,3,..,k). Hence the numerator determinant may be expressed as

(k) $$C_{11}\overline{C}_{11} - \sum_{i,j=2}^{k} C_{11}C_{1j}\overline{C}_{ij} .$$

But $|C_{ij}| = \overline{C}_{11}$(i,j=2,3,..,k). Dividing expression (k) by $\overline{C}_{11}$ and remembering that $\frac{\overline{C}_{ij}}{\overline{C}_{11}} = C^{ij}$(i,j = 2,3,...,k), we therefore establish the fact that $\bar{\sigma}^2_{1\cdot 23\ldots k}$ may be expressed as the ratio of determinants given in (h). The quantity $\bar{\sigma}^2_{1\cdot 23\ldots k}$ is the variance of X_1 about the least-square linear regression function (f), and should not be confused with $\sigma^2_{1\cdot 23\ldots k}$ as defined in §2.93.

The correlation coefficient between X_1 and the regression function (f) is known as the multiple correlation coefficient between X_1 and $X_2,X_3,..,X_k$ and is denoted by $R_{1\cdot 23\ldots k}$. To obtain an expression for the multiple correlation coefficient, we first determine the covariance between X_1 and the function (f), which is

(l) $$E[(X_1-a_1)(\sum_{i,j=2}^{k}(X_i-a_i)C_{ij}C^{ij})] = \sum_{i,j=2}^{k} C_{11}C_{ij}C^{ij}.$$

The variance of X_1 is C_{11} and that of (f) is

$$E[(\sum_{i,j=2}^{k}(X_i-a_i)C_{ij}C^{ij})^2],$$

whose value is equal to the last expression in (g), and which has been reduced to $\sum_{i,j=2}^{k} C_{11}C_{ij}C^{ij}$. Hence the multiple correlation coefficient is

(m) $$R_{1\cdot 23\ldots k} = \frac{\sum_{i,j=2}^{k} C_{11}C_{ij}C^{ij}}{\sqrt{C_{11}}\sqrt{\sum_{i,j=2}^{k} C_{11}C_{ij}C^{ij}}} = \sqrt{\frac{\sum_{i,j=2}^{k} C_{11}C_{ij}C^{ij}}{C_{11}}} .$$

It will be observed from (h) that

$$\bar{\sigma}^2_{1\cdot 23\ldots k} = C_{11}(1-R^2_{1\cdot 23\ldots k}),$$

and hence by §2.72, $R^2_{1\cdot 23\ldots k} = 1$ if, and only if, all of the probability in the k-dimensional space of the random variables lies on the least-square regression surface

$$x_1 - a_1 = \sum_{i,j=2}^{k} (x_i - a_i) C_{1j} C^{ij}.$$

It should be noted that a partial correlation coefficient between X_1 and X_2 with respect to X_r, X_{r+1}, ..., X_k could be determined for the case of a linear least-square regression function by replacing $a_{1\cdot x_r x_{r+1}\ldots x_k}$ and $a_{2\cdot x_r x_{r+1}\ldots x_k}$ by the corresponding linear least-square regression functions in determining $c_{12\cdot r(r+1)\ldots k}$, $\sigma_{1\cdot r(r+1)\ldots k}$, $\sigma_{2\cdot r(r+1)\ldots k}$ in §2.93.

Again, we remark that analogous results can be obtained by using an empirical c. d. f. $F_n(x_1, x_2, \ldots, x_k)$ instead of a probability c. d. f. $F(x_1, x_2, \ldots, x_k)$.

CHAPTER III

SOME SPECIAL DISTRIBUTIONS

In the present chapter, the notions of the preceding chapter will be exemplified by considering certain distributions that arise frequently in applied statistics. We shall begin by considering distributions for the discrete case. Since the distinction between the random variable X and the corresponding independent variable x of the distribution function has been made clear, we shall henceforth denote both by the lower case x unless this leads to ambiguity.

3.1 Discrete Distributions

3.11 Binomial Distribution

An important distribution function of a discrete variate is the binomial distribution which may be derived in the following manner. Suppose the probability of a "success" in a trial is p and the probability of a "failure" is $q = 1 - p$. For example the probability of a head in a toss of an "ideal" coin is $\frac{1}{2}$ and the probability of not a head (a tail) is $1 - \frac{1}{2} = \frac{1}{2}$. We can represent these probabilities in functional form $f(\alpha)$ where $f(\alpha) = p$ for $\alpha = 1$, a success, and $f(\alpha) = q$ for $\alpha = 0$, a failure. In other words $f(\alpha)$ is the probability of obtaining α successes in a single trial.

The probability associated with n trials which are mutually independent in the probability sense is

$$f(\alpha_1) \cdot f(\alpha_2) \cdot \ldots \cdot f(\alpha_n).$$

The probability of x successes and n - x failures in a specified order say $\alpha_1 = 1, \alpha_2 = 1, \ldots, \alpha_x = 1, \alpha_{x+1} = 0, \ldots, \alpha_n = 0$, is

$$f(1)^x f(0)^{n-x} = p^x q^{n-x}.$$

The number of orders in which x successes and n - x failures can occur is the number of combinations of n objects taken x at a time which is

(a) $$ {}_nC_x = \frac{n!}{x!(n-x)!}. $$

These ${}_nC_x$ orders are mutually exclusive events. Hence, to find the probability B(x), say, of exactly x successes irrespective of order we add the probabilities for all of the ${}_nC_x$ orders, thus obtaining

(b) $$B(x) = {}_nC_x p^x q^{n-x}.$$

B(x) will be recognized as the (x+1)-st term in the expansion of $(q+p)^n$. This demonstrates that the sum of the probabilities is equal to unity, i. e.

$$\sum_{x=0}^{n} B(x) = \sum_{x=0}^{n} {}_nC_x q^{n-x} p^x = (q+p)^n = 1^n = 1.$$

Hence $\sum_{x' \leq x} B(x')$ is clearly a c. d. f. F(x).

To derive the moments of the distribution B(x) we will find it convenient to use the m. g. f.

(c) $$\phi(\theta) = E(e^{x\theta}) = \sum_{x=0}^{n} {}_nC_x e^{\theta x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} {}_nC_x q^{n-x} (pe^{\theta})^x = (q+pe^{\theta})^n.$$

The h-th moment of x can be expressed as

$$E(x^h) = \mu_h' = \frac{\partial^h \phi(\theta)}{\partial \theta^h}\bigg|_{\theta = 0}$$

In particular the mean E(x) is

(d) $$\mu_1' = \frac{\partial}{\partial \theta} (q+pe^{\theta})^n \bigg|_{\theta = 0} = npe^{\theta}(q+pe^{\theta})^{n-1}\bigg|_{\theta = 0} = np,$$

and the second moment about zero is

$$\mu_2' = \frac{\partial^2}{\partial \theta^2} (q+pe^{\theta})^n \bigg|_{\theta = 0} = npe^{\theta}(q+pe^{\theta})^{n-1}\bigg|_{\theta = 0} + n(n-1)p^2e^{2\theta}(q+pe^{\theta})^{n-2}\bigg|_{\theta = 0}$$

$$= np + n(n-1)p^2.$$

Therefore, the variance is

(e) $$\sigma^2 = np + n(n-1)p^2 - n^2p^2 = np - np^2 = npq.$$

Example: Applying the binomial distribution to the coin tossing problem, we have $p = \frac{1}{2}$ and $q = \frac{1}{2}$. The probability of x heads is

$$B(x) = {}_nC_x(\tfrac{1}{2})^x(\tfrac{1}{2})^{n-x} = {}_nC_x(\tfrac{1}{2})^n$$

The mean and variance are, respectively,

$$\mu_1' = \frac{n}{2}, \qquad \sigma^2 = \frac{n}{4}.$$

In deducing B(x) we have assumed that p remains constant from trial to trial. If the probability is different for each trial, our conclusions must be modified. Let p_i be the probability of a success in the i-th trial (i = 1,2,...,n) and $q_i = 1 - p_i$ the corresponding probability of a failure. Let

$$p = \frac{1}{n}\sum_{i=1}^{n} p_i, \qquad q = \frac{1}{n}\sum_{i=1}^{n} q_i = 1 - p.$$

Then the expected value of $x = \sum_{i=1}^{n} \alpha_i$, the total number of successes in n trials, is

$$E(\alpha_1+\ldots+\alpha_n) = E(\alpha_1)+\ldots+E(\alpha_n) = p_1+\ldots+p_n = np.$$

The variance of α_i is p_iq_i. Since the trials are independent the variance of $x = \sum_{i=1}^{n} \alpha_i$ is $\sum_{i=1}^{n} p_iq_i$.

Noting that $p_i = p + (p_i-p)$ and $q_i = q - (p_i-p)$ we can write the variance

$$(f) \qquad \sum_{i=1}^{n} [p+(p_i-p)][q-(p_i-p)] = \sum_{i=1}^{n} [pq-(p_i-p)(p-q)-(p_i-p)^2] = npq - \sum_{i=1}^{n} (p_i-p)^2.$$

This is obviously less than the variance, npq, we found above. When the probability is constant from trial to trial, the distribution is known as the Bernoulli case; when the probability varies, we have the Poisson case.

In §2.71 it was proved that if a variate x is distributed about the mean a with the variance σ^2, we have the Tchebycheff inequality

$$Pr(|x-a|>\delta\sigma) \leq \frac{1}{\delta^2}$$

for any $\delta>0$. In the binomial distribution x has mean np and variance npq. Let us change to the variate $r = \frac{x}{n}$, the "relative frequency" of successes. We have $E(r) = E(\frac{x}{n}) = \frac{1}{n}E(x) = \frac{np}{n} = p$. Similarly, $\sigma_r^2 = \frac{npq}{n^2} = \frac{pq}{n}$. The Tchebycheff inequality states that

$$\Pr(|r-p|\geq\delta\sqrt{\frac{pq}{n}}) \leq \frac{1}{\delta^2}.$$

If we choose $\delta = \sqrt{\frac{n}{pq}}\lambda$, this inequality becomes

$$\Pr(|r-p|\geq\lambda) \leq \frac{pq}{n\lambda^2}. \tag{g}$$

Inequality (g) expresses what is known as the

Law of Large Numbers: For any given positive number λ, the probability that r will deviate from p by more than λ can be made arbitrarily small by choosing n sufficiently large.

Roughly speaking, the larger the value of n, the more the probability "piles up" around p (the mean of r) such that in the limit (as $n \longrightarrow \infty$) the probability is all piled at p.

In the example of "ideal" coin tossing r is the ratio of number of heads to total number of tosses. Then

$$\Pr(|r - \tfrac{1}{2}|\geq\lambda) \leq \frac{1}{4n\lambda^2}.$$

Example: If $\lambda = 0.1$ and $n = 100$, we have $\Pr(|r - \frac{1}{2}|\geq.1) < \frac{1}{400}$ In other words, the probability is less than $\frac{1}{400}$ that the relative frequency of heads will deviate from $\frac{1}{2}$ by more than 0.1.

3.12 Multinomial Distribution

An immediate generalization of the binomial distribution is the multinomial distribution. Suppose an event is characterized by a variate that can take on one and only one of k values, say $y_1, y_2, \ldots, y_k$. For example, if the event is the throw of a die and if y is the number of dots appearing on the top face, y can take on only one of the values 1, 2, 3, 4, 5, 6 in each throw. It should be noted that the k mutually exclusive kinds of events may not correspond to k values of a one-dimensional variable y. Thus, if $C_1, C_2, \ldots, C_k$ are k kinds of events, (e. g., the sides of a die may be colored rather than numbered) one and only one of which will occur in each trial, then we may let y be a vector with k components $(y^{(1)}, y^{(2)}, \ldots, y^{(k)})$, such that the value of the vector for an event of type C_1 is $(1,0,0,\ldots,0)$, the value for one of type C_2 is $(0,1,0,\ldots,0)$, etc. For convenience, we could denote these values of the vector y by y_1, y_2, etc., and proceed as in the case where $y_1, y_2, \ldots, y_k$ are different values of a one-dimensional variable y.

Let the probability of y being y_i be p_i where $\sum_{i=1}^{k} p_i = 1$. The probability associated with n trials is

$$f(y_{(1)}).f(y_{(2)})\ldots f(y_{(n)}),$$

where each of the y's will have one of the values $y_1, y_2,\ldots,y_k$, where $f(y_i) = p_i (i=1,2,\ldots,k)$. We now wish to find the probability that x_1 of the y's are y_1's, x_2 of the y's are y_2's, etc., $(\sum_{i=1}^{k} x_i = n)$.

The probability of x_1 events characterized by y_1, etc., occurring in a specified order, say $y_{(1)} = y_1,\ldots, y_{(x_1)} = y_1, y_{(x_1+1)} = y_2,\ldots, y_{(n)} = y_k$, is

$$f(y_1)^{x_1} f(y_2)^{x_2}\ldots f(y_k)^{x_k} = p_1^{x_1} p_2^{x_2}\ldots p_k^{x_k}.$$

The number of different orders in which we can get x_1 y_1's, etc., is the number of ways in which n objects can be permuted where x_1 are of type $C_1,\ldots,x_k$ are of type C_k, that is

$$\frac{n!}{x_1!x_2!\ldots x_k!}.$$

So the probability of x_1 y_1's, x_2 y_2's, etc., irrespective of the order in which they occur is given by adding the probabilities of various possible orders. We obtain

(a) $$M(x_1,x_2,\ldots,x_k) = \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2}\ldots p_k^{x_k}.$$

This may be recognized as the general term in the expansion of $(p_1+p_2+\ldots+p_k)^n$. Hence, the sum of $M(x_1,x_2,\ldots,x_k)$ over all partitions of n, that is, all sets of $x_i (\sum_{i=1}^{k} x_i = n, x_i \geq 0)$ is unity.

To find the means, variances, covariances, and higher moments we set up the m. g. f.

(b) $$\phi(\theta_1,\theta_2,\ldots,\theta_k) = E\left[e^{\sum_{1}^{k}\theta_i x_i}\right]$$

$$= \sum_{\sum x_i = n} \frac{n!}{x_1!x_2!\ldots x_k!} e^{\sum \theta_i x_i} p_1^{x_1} p_2^{x_2}\ldots p_k^{x_k}$$

$$= \sum_{\sum x_i = n} \frac{n!}{x_1!x_2!\ldots x_k!} (p_1 e^{\theta_1})^{x_1} (p_2 e^{\theta_2})^{x_2}\ldots (p_k e^{\theta_k})^{x_k}$$

$$= (p_1 e^{\theta_1} + \ldots + p_k e^{\theta_k})^n.$$

The mean of x_1 is

$$(c) \qquad E(x_1) = \frac{\partial \phi}{\partial \theta_1}\Big|_{\theta's=0} = np_1 e^{\theta_1}(p_1 e^{\theta_1}+\ldots+p_k e^{\theta_k})^{n-1}\Big|_{\theta's=0} = np_1.$$

And

$$E(x_1^2) = \frac{\partial^2 \phi}{\partial \theta_1^2}\Big|_{\theta's=0} = np_1 e^{\theta_1}(p_1 e^{\theta_1}+\ldots+p_k e^{\theta_k})^{n-1} + n(n-1)p_1^2 e^{2\theta_1}(p_1 e^{\theta_1}+\ldots+p_k e^{\theta_k})^{n-2}\Big|_{\theta's=0}$$

$$= np_1 + n(n-1)p_1^2.$$

Therefore, the variance of x_1 is

$$(d) \qquad \sigma_{x_1}^2 = np_1 + n(n-1)p_1^2 - n^2 p_1^2 = np_1(1-p_1).$$

In a similar manner we find the covariance between x_1 and x_j to be $-np_1p_j$. It is clear that the binomial distribution is the special case when $k = 2$.

3.13 The Poisson Distribution

The Poisson distribution is in a sense a particular limiting form of the binomial distribution. We shall deduce it from geometrical considerations. Let AB be a line segment of length L and CD a segment of length l contained in AB.

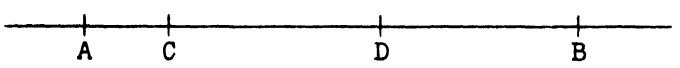

Figure 3

Let the probability that a point taken at random falls on an interval of length du be $\frac{du}{L}$; that is, the p. d. f. of u is a constant. The probability of the point falling in CD is $\frac{l}{L}$. If we let n points fall at random on AB, the probability that exactly x of them fall on CD is given by the binomial distribution ((a) of §3.11)

$$B(x) = \frac{n!}{x!(n-x)!}\left(\frac{l}{L}\right)^x\left(1-\frac{l}{L}\right)^{n-x}.$$

Now let n and L increase indefinitely in such a way that the average number of points per unit length is a finite number $k \neq 0$, i. e., $\frac{n}{L} \longrightarrow k$. Now

$$B(x) = \frac{n(n-1)\ldots(n-x+1)}{[x!n^x]}\left(\frac{nl}{L}\right)^x\left(1-\frac{n}{L}\frac{l}{n}\right)^{n-x}.$$

So the limiting value of B(x) for a given x is

$$\lim_{n,L\to\infty} B(x) = \lim_{n,L\to\infty} \frac{1(1-\frac{1}{n})\ldots(1-\frac{x+1}{n})}{x!} (\frac{n}{L}l)^x (1-\frac{n}{L}\cdot\frac{l}{n})^{n-x} = \frac{(kl)^x e^{-kl}}{x!} .$$

Let kl = m, and we get the usual expression for the Poisson distribution

(a) $$p(x) = \frac{e^{-m}m^x}{x!} .$$

The sum over all x is seen to be 1,

$$\sum_{x=0}^{\infty} \frac{e^{-m}m^x}{x!} = e^{-m}(1+m+\frac{m^2}{2!}+\ldots) = e^{-m}e^m = 1.$$

The m. g. f. is

(b) $$\phi(\theta) = e^{-m}\sum_{x=0}^{\infty} e^{\theta x}\frac{m^x}{x!} = e^{-m}\sum_{x=0}^{\infty}\frac{(me^\theta)^x}{x!} = e^{-m}.e^{me^\theta} = e^{m(e^\theta-1)} .$$

From this we derive the moments about zero in the customary manner.

$$E(x) = \frac{\partial\phi}{\partial\theta}\bigg|_{\theta=0} = me^\theta e^{m(e^\theta-1)}\bigg|_{\theta=0} = m,$$

(c)
$$E(x^2) = \frac{\partial^2\phi}{\partial\theta^2}\bigg|_{\theta=0} = me^\theta e^{m(e^\theta-1)} + m^2e^{2\theta}e^{m(e^\theta-1)}\bigg|_{\theta=0} = m + m^2.$$

Therefore, the variance is equal to the mean.

(d) $$\sigma^2 = m + m^2 - m^2 = m.$$

This argument given for one dimension immediately extends to two or more dimensions. For example, for two dimensions we would take AB and CD to be regions of the plane, the latter contained in the former, and k to be the limiting ratio of the number of points per unit area. The Poisson distribution is applicable to problems dealing with occurence of events in a time interval of a given length such as emission of rays from radioactive substances, certain traffic problems, demands for telephone service and bacteric count in cells.

Example: Let us consider the following problem as an example to which the Poisson distribution is applicable. If X-rays are considered as discrete quanta and if the absorption of k or more will kill a certain unicellular organism, what is the probability that an organism of a given size S on a given glass slide will escape death by X-rays after being exposed for t seconds? On the assumption that the

projection of the organism of size S on a plane has an area of a, and m is the average number of rays striking an area of size a in t seconds, and the rays appear independently and at random, then the probability that x of the X-rays hit the organism in t seconds is

$$p(x) = \frac{e^{-m}m^x}{x!} .$$

Hence, the probability of survival is $\sum_{x=0}^{k-1} p(x)$. The average number of rays absorbed by the survivors is

$$\sum_{x=0}^{k-1} xp(x) \Big/ \sum_{x=0}^{k-1} p(x).$$

3.14 The Negative Binomial Distribution

Another discrete distribution which is closely related to the Bernoulli binomial distribution is the negative binomial. If we expand, according to the binomial theorem,

$$(q - p)^{-k}$$

where $q = 1 + p$, $k > 0$, $p > 0$, we get as the general term

(a) $$q^{-k} \frac{\Gamma(k+x)}{x!\Gamma(k)} \left(\frac{p}{q}\right)^x.$$

When we interpret this as a probability function of x, p(x), it is called the negative binomial distribution and is defined for $x = 0, 1, 2, \ldots$ We notice that the sum of p(x) for all x is unity,

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} q^{-k} \frac{\Gamma(k+x)}{x!\Gamma(k)} \left(\frac{p}{q}\right)^x = (q-p)^{-k} = 1^{-k} = 1.$$

The m. g. f. is

(b) $$\phi(\theta) = \sum_{x=0}^{\infty} q^{-k} e^{\theta x} \frac{\Gamma(k+x)}{x!\Gamma(k)} \left(\frac{p}{q}\right)^x = \sum_{x=0}^{\infty} q^{-k} \frac{\Gamma(k+x)}{x!\Gamma(k)} \left(\frac{pe^{\theta}}{q}\right)^x = (q-pe^{\theta})^{-k}.$$

From this we find the mean

(c) $$E(x) = \frac{\partial \phi}{\partial \theta}\Big|_{\theta=0} = kpe^{\theta}(q-pe^{\theta})^{-(k+1)}\Big|_{\theta=0} = kp$$

and

$$E(x^2) = \frac{\partial^2 \phi}{\partial \theta^2}\bigg|_{\theta=0} = kpe^{\theta}(q-pe^{\theta})^{-(k+1)} + k(k+1)p^2e^{2\theta}(q-pe^{\theta})^{-(k+2)}\bigg|_{\theta=0} = kp + k(k+1)p^2.$$

Therefore the variance is

$$\sigma^2 = kp + k(k+1)p^2 - k^2p^2 = kp + kp^2 = kpq. \tag{d}$$

The similarity of this m. g. f. and these moments to those of the positive binomial distribution should be noted.

It can easily be shown that a special limiting case of the negative binomial distribution is the Poisson law. If we let $p \longrightarrow 0$ and $k \longrightarrow \infty$ in such a way that

$$\lim kp = m,$$

then

$$\lim_{\substack{k\to\infty \\ p\to 0}} (1+p)^{-k} \frac{\Gamma(k+x)}{x!\Gamma(k)} \left(\frac{p}{1+p}\right)^x$$

$$= \lim_{k\to\infty} \left(1+\frac{m}{k}\right)^{-k} \frac{1}{x!}\cdot(k+x)(k+x-1)\ldots(k+1)\left(\frac{m}{k}\right)^x\left(1+\frac{m}{k}\right)^{-x}$$

$$= \lim_{k\to\infty} \left(1+\frac{m}{k}\right)^{-k} \frac{1}{x!}\left(1+\frac{x}{k}\right)\left(1+\frac{x-1}{k}\right)\ldots\left(1+\frac{1}{k}\right) m^x \left(1+\frac{m}{k}\right)^{-x}$$

$$= e^{-m}\frac{m^x}{x!}.$$

If we make a change of parameters, we have the usual expression for the Polya-Eggenburger distribution. Let

$$k = \frac{h}{d}, \qquad p = d.$$

Then the distribution may be written as

$$p(x) = (1+d)^{-\frac{d}{h}} \frac{\Gamma(\frac{h}{d}+x)}{x!\Gamma(\frac{h}{d})} \left(\frac{d}{1+d}\right)^x. \tag{e}$$

This distribution, one of a number of contagious distributions, is useful in describing, for example, the probability of x cases of a given epidemic in a given locality.

If we interpret $\frac{1}{q}$ as the probability of a "success" and $\frac{p}{q}$ as the probability of a "failure" in a trial, then it will be seen that (a) is the probability that x + k trials will be required to obtain k successes. For the probability of obtaining k - 1 successes

and x failures in x + k - 1 trials is

$$\frac{(x+k-1)!}{(k-1)!\,x!}\left(\frac{p}{q}\right)^{x}\left(\frac{1}{q}\right)^{k-1}.$$

Now the last trial must be a success. Therefore multiplying this probability by $(\frac{1}{q})$, the probability of success, we obtain (a), the probability that x + k trials will be required to obtain k successes.

3.2 The Normal Distribution

3.21 The Univariate Case. A very important distribution is the normal or Gaussian distribution

$$ke^{-h^2(x-c)^2}$$

defined over the range $-\infty < x < \infty$ where k, h, and c are constants. Various attempts have been made to establish this distribution from postulates and other primitive assumptions. Gauss, for example, deduced it from the postulate of the arithmetic mean which states, roughly, that for a set of equally valid observations of a quantity the arithmetic mean is the most probable value. Pearson derived it as a solution of a certain differential equation. It can be shown that it is the limiting distribution of the Bernoulli binomial distribution. We shall not derive the normal distribution from more basic considerations, but we shall observe that it arises under rather broad conditions as a limiting distribution in many situations involving a large number of variates.

We can determine k in the distribution by requiring that the integral over the entire range be unity. If we let $u = h(x-c)$, we wish

$$\int dF(x) = \frac{k}{h}\int_{-\infty}^{\infty} e^{-u^2}du = 1.$$

To evaluate the integral $I = \int_{-\infty}^{\infty} e^{-u^2}du$ we observe that

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-u^2-v^2}dudv.$$

Changing to polar coordinates $u = r\cos\theta$, $v = r\sin\theta$, we get

$$I^2 = \int_0^{2\pi}\int_0^{\infty} re^{-r^2}dr\,d\theta = \int_0^{2}\frac{d\theta}{2} = \pi.$$

Therefore, we take $k = \frac{\sqrt{h}}{\sqrt{\pi}}$.

The mean of the distribution is

$$E(x) = \sqrt{\frac{h}{\pi}} \int_{-\infty}^{\infty} x e^{-h^2(x-c)^2} dx$$

$$= \sqrt{\frac{h}{\pi}} \int_{-\infty}^{\infty} c e^{-h^2(x-c)^2} dx + \sqrt{\frac{h}{\pi}} \int_{-\infty}^{\infty} (x-c) e^{-h^2(x-c)^2} dx.$$

The latter integral is zero because the integrand is an odd function of $x - c$. So

$$a = E(x) = c\sqrt{\frac{h}{\pi}} \int_{-\infty}^{\infty} e^{-h^2(x-c)^2} dx = c.$$

The variance is found by integration by parts,

$$\sigma^2 = E[(x-c)^2] = \sqrt{\frac{h}{\pi}} \int_{-\infty}^{\infty} (x-c)^2 e^{-h^2(x-c)^2} dx = \frac{1}{2h^2}.$$

We usually write the normal distribution with c and h expressed in terms of a and σ^2, respectively, i. e.,

(a) $$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

We shall refer to this distribution as $N(a,\sigma^2)$.

To find higher moments (about the mean) it is convenient to use the m. g. f. of the normalized variate $\frac{x-a}{\sigma}$.

$$\phi(\theta) = E(e^{\theta(\frac{x-a}{\sigma})}) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\theta\frac{x-a}{\sigma}} e^{-\frac{(x-a)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[\frac{x-a}{\sigma} - \theta]^2 + \frac{1}{2}\theta^2} dx.$$

Setting $\frac{x-a}{\sigma} - \theta = y$, the last integral becomes

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2 + \frac{1}{2}\theta^2} dy = e^{\frac{1}{2}\theta^2}$$

Hence,

(b) $$\phi(\theta) = e^{\frac{1}{2}\theta^2}$$

It should be noticed that the normal distribution is symmetrical with respect to the line $x = a$, its mean. The smaller the value of σ^2 is, the greater the concentration about the mean. In fact σ is the distance from the mean to the points of inflection:

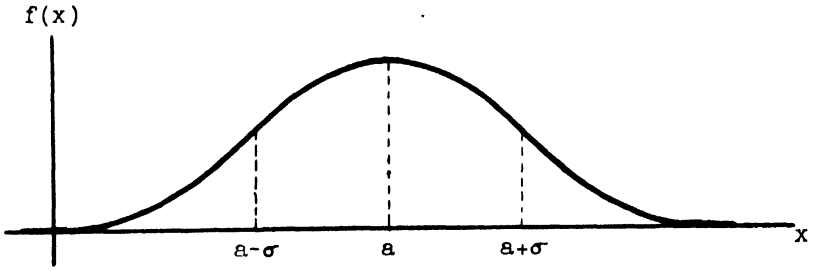

Figure 4

Because of its wide application and because of its theoretical importance, the normal distribution has been the origin of much of the terminology and many of the concepts in statistics.

The integral

$$\frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du = 1 - F(x)$$

is widely tabulated; the ordinate

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

is also tabulated in many places. The value of x for which

$$\frac{1}{\sqrt{2\pi}} \int_{-x}^{+x} e^{\frac{-u^2}{2}} du = \frac{1}{2}$$

is called the probable error and is approximately .6745.

It can be readily verified by applying Theorem (C) §4.21, that as $n \to \infty$ the normalized variable $\frac{x-np}{\sqrt{npq}}$, where x is distributed according to the binomial law, has the limiting distribution $N(0,1)$. For we may write

$$\frac{x-np}{\sqrt{npq}} = \frac{(\sum_1^n \frac{x_i}{n} - p)\sqrt{n}}{\sqrt{pq}}$$

where $x_1, x_2, \ldots, x_n$ are independently distributed according to the law $p(x) = p^x(1-p)^{1-x}$, $(x=0, \text{ or } 1)$. The mean of this distribution is $E(x) = \sum_{x=0}^{1} xp^x(1-p)^{1-x} = p$, and the variance is $\sigma^2 = \sum_{x=0}^{1} (x-p)^2 p^x (1-p)^{1-x} = pq$. The applicability of Theorem (C), §4.21, is then obvious.

3.22 The Normal Bivariate Distribution

The extension of the normal probability density function to the case of two variables, x_1 and x_2, is straight forward. We replace $(x-a)^2$ by a quadratic form in $x_1 - a_1$ and $x_2 - a_2$. The distribution may be written

$$Ke^{-\frac{1}{2}Q}$$

where $Q = A_{11}y_1^2 + 2A_{12}y_1y_2 + A_{22}y_2^2$, $y_1 = x_1 - a_1$, and K, $A_{11} > 0$, $A_{22} > 0$, A_{12} are constants such that $A_{11}A_{22} > A_{12}^2$. These inequalities on the A's are necessary and sufficient conditions for Q to be a positive definite quadratic form in y_1 and y_2, i. e., $Q > 0$ unless $y_1 = y_2 = 0$. We wish to determine K so that the integral of the p. d. f. over the x_1x_2-plane is unity. The integral transforms to

$$\text{(a)} \qquad \begin{aligned} &K\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(A_{11}y_1^2+A_{22}y_2^2+2A_{12}y_1y_2)}\,dy_1dy_2 \\ &= K\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}A_{11}(y_1^2+2\frac{A_{12}}{A_{11}}y_1y_2+\frac{A_{12}^2}{A_{11}^2}y_2^2)-\frac{1}{2}(A_{22}-\frac{A_{12}^2}{A_{11}})y_2^2}\,dy_1dy_2 . \end{aligned}$$

If we let $y_1 + \frac{A_{12}}{A_{11}}y_2 = z_1$, and integrate z_1 and y_2 in (a) from $-\infty$ to $+\infty$, and use the fact that

$$\int_{-\infty}^{\infty} e^{-cx^2}dx = \sqrt{\frac{\pi}{c}} \qquad (c>0),$$

we obtain for (a)

$$K \cdot \frac{2\pi}{\sqrt{A_{11}A_{22}-A_{12}^2}} .$$

If the integral is to be unity, we must choose

$$K = \frac{\sqrt{A_{11}A_{22}-A_{12}^2}}{2\pi} = \frac{\sqrt{A}}{2\pi} ,$$

where A is the determinant

$$\begin{vmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{vmatrix} .$$

We may, therefore, write the distribution as

(b) $$\frac{\sqrt{A}}{2\pi} e^{-\frac{1}{2}\sum_{i,j=1}^{2} A_{ij}(x_i-a_i)(x_j-a_j)},$$

where $A_{ij} = A_{ji}$.

In order to find the means, variances, and covariance of x_1 and x_2, it will be convenient to obtain the m. g. f. of (x_1-a_1) and (x_2-a_2), i. e.

(c) $$\phi(\theta_1,\theta_2) = E\left(e^{\sum_{i=1}^{2}\theta_i(x_i-a_i)}\right) = \frac{\sqrt{A}}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}Q+\sum_{i=1}^{2}\theta_i(x_i-a_i)}\,dx_1dx_2.$$

Letting $x_i - a_i = y_i$, we have

(d) $$\phi(\theta_1,\theta_2) = \frac{\sqrt{A}}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{i,j=1}^{2}A_{ij}y_iy_j+\sum_{i}^{2}\theta_iy_i}\,dy_1dy_2$$
$$= \frac{\sqrt{A}}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}A_{11}\left(y_1+\frac{A_{12}}{A_{11}}y_2-\frac{\theta_1}{A_{11}}\right)^2-\frac{1}{2}\left(A_{22}-\frac{A_{12}^2}{A_{11}}\right)\left(y_2+\frac{A_{12}\theta_1-A_{11}\theta_2}{A}\right)^2+\frac{1}{2}R}\,dy_1dy_2$$

where $R = \frac{A_{22}\theta_1^2+A_{11}\theta_2^2-2A_{12}\theta_1\theta_2}{A} = A^{11}\theta_1^2+A^{22}\theta_2^2+2A^{12}\theta_1\theta_2$ where $A^{ij} = \frac{\text{cofactor of } A_{ij} \text{ in } A}{A}$.

Making the change of variables

$$y_1 + \frac{A_{12}}{A_{11}}y_2 - \frac{\theta_1}{A_{11}} = z_1,\quad y_2 + \frac{A_{12}\theta_1-A_{11}\theta_2}{A} = z_2,$$

and integrating with respect to z_1 and z_2, we obtain

(e) $$\phi(\theta_1,\theta_2) = e^{\frac{1}{2}\sum_{i,j=1}^{2}A^{ij}\theta_i\theta_j}.$$

Now consider the problem of finding the mean values of x_1 and x_2. We have

$$E(x_1-a_1) = \frac{\partial\phi}{\partial\theta_1}\Big|_{\theta_1=\theta_2=0} = (\theta_1A^{11}+\theta_2A^{12})\phi(\theta_1,\theta_2)\Big|_{\theta_1=\theta_2=0} = 0.$$

Hence $E(x_1) = a_1$. Similarly $E(x_2) = a_2$.

To find the variances and covariances of x_1 and x_2, we must take second derivatives. Thus to find the variance of x_1 we have

$$\sigma_1^2 = E[(x_1-a_1)^2] = \frac{\partial^2\phi}{\partial\theta_1^2}\Bigg|_{\theta_1=\theta_2=0} = [A^{11}+(\theta_1A^{11}+\theta_2A^{12})^2]\phi(\theta_1,\theta_2)\Big|_{\theta_1=\theta_2=0} = A^{11}.$$

Similarly,

$$\sigma_2^2 = A^{22}.$$

For the covariance, we have

$$\rho\sigma_1\sigma_2 = E[(x_1-a_1)(x_2-a_2)] = \frac{\partial^2\phi}{\partial\theta_1\partial\theta_2}\Bigg|_{\theta_1=\theta_2=0}$$

$$= [A^{12}+(\theta_1A^{11}+\theta_2A^{12})(\theta_2A^{22}+\theta_1A^{12})]\phi(\theta_1,\theta_2)\Big|_{\theta_1=\theta_2=0} = A^{12}.$$

If the three equations

$$\sigma_1^2 = A^{11}$$

(f) $$\sigma_2^2 = A^{22}$$

$$\sigma_1\sigma_2\rho = A^{12}$$

are solved for A_{11}, A_{12}, A_{22}, we obtain

(g) $$A_{11} = \frac{1}{\sigma_1^2(1-\rho^2)}, \quad A_{22} = \frac{1}{\sigma_2^2(1-\rho^2)}, \quad A_{12} = \frac{-\rho}{\sigma_1\sigma_2(1-\rho^2)}.$$

We may summarize as follows:

Theorem (A): If x_1, x_2 are distributed according to the bivariate normal distribution

(h) $$\frac{\sqrt{A}}{2\pi}\, e^{-\frac{1}{2}\sum_{i,j=1}^{2} A_{ij}(x_i-a_i)(x_j-a_j)},$$

the m. g. f. of (x_1-a_1) and (x_2-a_2) is given by (e); $E(x_i) = a_i$, $(i=1,2)$; the variance of x_i is A^{ii} $(i=1,2)$ and the covariance between x_1 and x_2 is A^{12}. A_{11}, A_{22}, A_{12} are expressed in terms of variances and the correlation coefficient between x_1 and x_2 by (g).

Expressing A_{11}, A_{12}, A_{22} in (h) in terms of σ_1^2 σ_2^2 and ρ, the distribution (h) may be written as

$$(1)\qquad \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-a_1)^2}{\sigma_1^2}+\frac{(x_2-a_2)^2}{\sigma_2^2}-2\rho\frac{(x_1-a_1)(x_2-a_2)}{\sigma_1\sigma_2}\right]}.$$

The marginal distribution of (i) with respect to x_1 is the distribution of x_1. Thus integrating (i) with respect to x_2 we obtain as the distribution of x_1

$$f_1(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1}\, e^{-\frac{1}{2}\frac{(x_1-a_1)^2}{\sigma_1^2}}.$$

A similar expression holds for the distribution of x_2.

We would also like to know the conditional probability function

$$f(x_2|x_1) = \frac{f(x_1,x_2)}{f_1(x_1)}.$$

Substituting the expressions for $f(x_1,x_2)$ and $f_1(x_1)$ from (a) and (b), respectively, we find

$$(j)\qquad f(x_2|x_1) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\frac{[x_2-a_2-\rho\frac{\sigma_2}{\sigma_1}(x_1-a_1)]^2}{\sigma_2^2(1-\rho^2)}}.$$

Thus, for a fixed value of x_1, x_2 is distributed according to $N(a_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-a_1),\ \sigma_2^2(1-\rho^2))$.

In a similar way we can show that the marginal distribution of x_2 is $N(a_2,\sigma_2^2)$ and the conditional probability of x_1, given x_2, is $N(a_1+\rho\frac{\sigma_1}{\sigma_2}(x_2-a_2),\sigma_1^2(1-\rho^2))$. It will be observed that if $\rho = 0$, the marginal and the conditional probability distributions of x_1 (or x_2) are identical.

Since the conditional distribution of x_2 is $N(a_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-a_1),\ \sigma_2^2(1-\rho^2))$, the mean value of x_2 for the interval (x_1,x_1+dx_1) is simply $a_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-a_1)$. So the regression function of x_2 on x_1 is linear, that is,

$$(k)\qquad a_{2.x_1} = a_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-a_1).$$

Similarly

$$(l)\qquad a_{1.x_2} = a_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-a_2).$$

Since $\sigma_2^2(1-\rho^2)$ is the variance of x_2 about the mean $a_{2.x_1}$ in the conditional probability distribution, the nearer ρ^2 is to 1, the smaller is this variance. If $\rho = 0$, x_2 does not

depend on x_1; the two variates are independent and

$$f(x_1,x_2) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-a_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_2-a_2)^2}{2\sigma_2^2}}.$$

3.23 The Normal Multivariate Distribution

Let us now consider the extension of §3.22 to the case of k variates.

Let

$$f(x_1,x_2,\ldots,x_k) = Ce^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}(x_i-a_i)(x_j-a_j)},$$

where $\| A_{ij} \|$ is a symmetric, positive definite matrix, that is, $A_{ij} = A_{ji}$ and $\sum_{i,j=1}^{k} A_{ij}t_it_j > 0$ for real t_i, not all zero.

We wish to determine C so that the integral over the entire range, $-\infty < x_i < \infty$, is unity. We must have

$$\frac{1}{C} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum A_{ij}(x_i-a_i)(x_j-a_j)} dx_1 \ldots dx_k.$$

To evaluate this integral, we transform the variables. Let

$$x_1 - a_1 = y_1.$$

Then

$$\frac{1}{C} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}Q} dy_1 dy_2 \ldots dy_k,$$

where $Q = \sum_{i,j=1}^{k} A_{ij}y_iy_j$.

Now we can write

$$Q = A_{11}\left[y_1 + \frac{\sum_{j=2}^{k} A_{1j}y_j}{A_{11}}\right]^2 + \sum_{i,j=2}^{k} \left(A_{ij} - \frac{A_{11}A_{1j}}{A_{11}}\right) y_iy_j.$$

Let

$$z_1 = y_1 + \frac{\sum_{j=2}^{k} A_{1j}y_j}{A_{11}},$$

$$A_{ij}^{(1)} = A_{ij} - \frac{A_{1j}A_{i1}}{A_{11}}, \qquad i,j = 2,\ldots,k.$$

Then

$$\frac{1}{C} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}[A_{11}z_1^2 + \sum_{i,j=2}^{k} A_{ij}^{(1)} y_i y_j]} dz_1 dy_2 \cdots dy_k.$$

The range of z_1 is $-\infty < z_1 < \infty$.

We should observe that the quadratic form is again positive definite, that is,

$$A_{11}s_1^2 + \sum_{i,j=2}^{k} A_{ij}^{(1)} s_i s_j > 0$$

for real s_i not all zero. For if there were such a set of s's for which this quadratic form were zero or negative, it would be implied that there is a set of t's for which $\sum_{i,j=1}^{k} A_{ij} t_i t_j \leq 0$.

We continue this process, in turn letting

$$z_2 = y_2 + \frac{\sum_{j=2}^{k} A_{2j}^{(1)} y_j}{A_{22}^{(1)}},$$

$$\vdots$$

$$z_k = y_k,$$

and correspondingly

$$A_{ij}^{(2)} = A_{ij}^{(1)} - \frac{A_{2j}^{(1)} A_{i2}^{(1)}}{A_{22}^{(1)}}, \qquad i,j = 3,\ldots,k,$$

$$\vdots$$

$$A_{kk}^{(k-1)} = A_{kk}^{(k-2)} - \frac{A_{k-1,k}^{(k-2)} A_{k,k-1}^{(k-2)}}{A_{k-1,k-1}^{(k-2)}}.$$

Each quadratic form in this sequence is positive definite by the foregoing argument. The integral becomes

$$\frac{1}{C} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}A_{11}z_1^2 - \frac{1}{2}A_{22}^{(1)}z_2^2 - \cdots - \frac{1}{2}A_{kk}^{(k-1)}z_k^2} dz_1 \cdots dz_k.$$

The final quadratic form is positive definite, so $A_{11} > 0$, $A_{22}^{(1)} > 0$, $\ldots, A_{kk}^{(k-1)} > 0$. Hence we can integrate on each z in turn, using the fact that

$$\int_{-\infty}^{\infty} e^{-cx^2} = \frac{\sqrt{\pi}}{\sqrt{c}}.$$

Therefore, we get

$$\frac{1}{C} = \frac{(2\pi)^{\frac{k}{2}}}{\sqrt{A_{11}A_{22}^{(1)}\ldots A_{kk}^{(k-1)}}}.$$

To find the value of $\frac{1}{C}$, let us evaluate by Lagranges' method (known also as pivotal condensation) the determinant of $\| A_{ij} \|$

$$|A| = \begin{vmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ A_{21} & A_{22} & \cdots & A_{2k} \\ \vdots & & & \\ A_{k1} & A_{k2} & \cdots & A_{kk} \end{vmatrix} = A_{11} \begin{vmatrix} 1 & A_{12} & \cdots & A_{1k} \\ \frac{A_{21}}{A_{11}} & A_{22} & \cdots & A_{2k} \\ \vdots & & & \\ \frac{A_{k1}}{A_{11}} & A_{k2} & \cdots & A_{kk} \end{vmatrix}.$$

If we subtract A_{12} times the first column from the second, etc., we get

$$|A| = A_{11} \begin{vmatrix} 1 & 0 & \cdots & 0 \\ \frac{A_{21}}{A_{11}} & A_{22} - \frac{A_{21}A_{12}}{A_{11}} & \cdots & A_{2k} - \frac{A_{21}A_{1k}}{A_{11}} \\ \vdots & & & \\ \frac{A_{k1}}{A_{11}} & \frac{A_{k1}A_{12}}{A_{11}} & \cdots & A_{kk} - \frac{A_{k1}A_{1k}}{A_{11}} \end{vmatrix}.$$

$$= A_{11} \begin{vmatrix} A_{22}^{(1)} & \cdots & A_{2k}^{(1)} \\ \vdots & & \\ A_{k2}^{(1)} & \cdots & A_{kk}^{(1)} \end{vmatrix}$$

Continuing in this way, we find the value of the determinant

(a) $$|A| = A_{11}A_{22}^{(1)}\ldots A_{kk}^{(k-1)}.$$

Therefore, the constant we are seeking is

$$C = \frac{\sqrt{|A|}}{(2\pi)^{\frac{k}{2}}},$$

and the normal multivariate p. d. f. is

(b) $$\frac{\sqrt{|A|}}{(2\pi)^{\frac{k}{2}}} e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}(x_i-a_i)(x_j-a_j)}.$$

At this point we should notice some properties of positive definite quadratic forms and matrices. Since $|A| = A_{11}A_{22}^{(1)} \dots A_{kk}^{(k-1)}$, $|A|$ is positive, for each of the factors is a positive constant. Corresponding to each principal minor of $\| A_{ij} \|$ of order h, there is a quadratic form in h variables. This quadratic form is again positive definite. For if there were a set of h t's (not all zero) making this form zero or negative, this set and the (k-h) other t's zero would do the same for $\sum_{i,j=1}^{k} A_{ij}t_it_j$. Since the determinant of a positive definite matrix is positive, it follows that every principal minor is positive. Conversely, if every principal minor is positive the matrix or the quadratic form is positive definite, for then each $A_{ii}^{(j)}$ is positive and the above process of reducing to a sum of squares may be carried out.

The transformation to the z's is linear, of the form

$$z_i = \sum_{j=1}^{k} b_{ij}x_j,$$

where $b_{ij} = 0$ for $j < i$. The process we have used proves the theorem that any positive definite quadratic form may be "diagonalized" by a real linear transformation. If we followed this by the transformation

$$w_i = \sqrt{A_{ii}^{(i-1)}}z_i$$

we would have reduced the quadratic form to a sum of squares. This last is equivalent to

$$\text{(c)} \qquad \begin{aligned} w_i &= \sqrt{A_{ii}^{(i-1)}}\sum_{j=1}^{k} b_{ij}x_j \\ &= \sum_{j=1}^{k} C_{ij}x_j, \end{aligned}$$

where $C_{ij} = \sqrt{A_{ii}^{(i-1)}}b_{ij}$.

Now we wish to show that the mean is

$$E(x_i) = a_i.$$

To do this we differentiate both sides of the following equation with respect to a_i:

$$\frac{\sqrt{|A|}}{(2\pi)^{k/2}} \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum A_{ij}(x_i-a_i)(x_j-a_j)} dx_1 \dots dx_k = 1.$$

Since $\frac{\partial}{\partial a_h} \sum_{i,j} A_{ij}(x_i-a_i)(x_j-a_j) = -2 \sum_{j=1}^{k} A_{hj}(x_j-a_j)$ the differentiation of the above equation gives us

$$\frac{\sqrt{|A|}}{(2\pi)^{k/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [\sum_j A_{ij}(x_j-a_j)] e^{-\frac{1}{2}\sum A_{ij}(x_i-a_i)(x_j-a_j)} dx_1 \ldots dx_k = 0.$$

So

$$E[\sum_j A_{ij}(x_j-a_j)] = \sum_j A_{ij}E(x_j-a_j) = 0$$

for $i = 1, 2, \ldots, k$. This gives us k homogeneous linear equations in the k unknowns, $E(x_j-a_j)$. Since the determinant of the coefficient matrix, $|A|$, is not equal to zero, the only solution to these equations is that all the unknowns be zero.

$$E(x_j-a_j) = 0.$$

So

(d) $$E(x_j) = a_j, \qquad j = 1,2,\ldots,k.$$

Next we wish to show that the covariance of x_i and x_j is

$$E[(x_i-a_i)(x_j-a_j)] = A^{ij} = \frac{\text{cofactor of } A_{ij} \text{ in } \| A_{ij} \|}{|A|}.$$

To demonstrate this we differentiate with respect to A_{ij} both sides of the identity

$$\frac{1}{(2\pi)^{k/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum A_{ij}(x_i-a_i)(x_j-a_j)} dx_1 \ldots dx_k = |A|^{-\frac{1}{2}}.$$

Differentiating, we have

$$\frac{1}{(2\pi)^{k/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [(\frac{\delta_{ij}-2}{2})(x_i-a_i)(x_j-a_j)] e^{-\frac{1}{2}\sum A_{ij}(x_i-a_i)(x_j-a_j)} dx_1 \ldots dx_k$$

$$= (\frac{\delta_{ij}-2}{2})|A|^{-\frac{3}{2}} \cdot (\text{cofactor of } A_{ij}) = (\frac{\delta_{ij}-2}{2})|A|^{-\frac{1}{2}} A^{ij},$$

where $\delta_{ij} = 1$ if $i = j$, and $= 0$ if $i \neq j$.

If we multiply both sides of this equation by $(\frac{2}{\delta_{ij}-2})\,|A|^{\frac{1}{2}}$ the left hand side is

$$E[(x_i-a_i)(x_j-a_j)],$$

and the right hand side is A^{ij}. So we have

(e) $$\sigma_1^2 = E[(x_1-a_1)^2] = A^{11}, \qquad i = 1,2,..,k,$$

(f) $$\sigma_i\sigma_j\rho_{ij} = E[(x_i-a_i)(x_j-a_j)] = A^{ij}, \qquad i\neq j,\ i,j = 1,2,\ldots,k.$$

We may summarize as follows:

Theorem (A): If $x_1, x_2,\ldots, x_k$ are distributed according to the normal multivariate distribution (b), then $E(x_i) = a_i$, $\sigma_i^2 = A^{ii}$, and $\sigma_i\sigma_j\rho_{ij} = A^{ij}$.

Now let us find the joint marginal distribution of $x_1, x_2,..,x_r$ $(r<k)$. To do this we integrate out $x_{r+1},\ldots,x_k$, getting

(g) $$g(x_1,\ldots,x_r) = \frac{\sqrt{|B|}}{(2\pi)^{r/2}}\, e^{-\frac{1}{2}\sum_{u,v=1}^{r} B_{uv}(x_u-a_u)(x_v-a_v)}.$$

We can see this is true if we recall the procedure used in evaluating $\frac{1}{C}$. If at any stage, we had integrated out the z's, we would have had remaining a normal multivariate distribution of the x's.

We wish to find an expression of the B_{uv} in terms of the A_{ij}. We know that the value of $E[(x_u-a_u)(x_v-a_v)]$ is A^{uv} if found from the original distribution and is B^{uv} if found from the marginal distribution. But these two expressions must be equal. Therefore

$$A^{uv} = B^{uv}.$$

Hence, to derive $||B_{uv}||$ from $||A_{ij}||$ we delete from $||A^{ij}||$ the last k - r rows and columns (obtaining $||B^{uv}||$) and take the inverse of this matrix.

In particular, suppose r = 1. We find the distribution of x_1 to be

$$g(x_1) = \frac{\sqrt{B_{11}}}{\sqrt{2\pi}}\, e^{-\frac{1}{2}B_{11}(x_1-a_1)^2},$$

where

$$B_{11} = |A^{11}|^{-1} = \frac{|A|}{\bar{A}_{11}},$$

where $\bar{A}_{11}$ = cofactor of A_{11} in A.

Thus, $$\sigma_1^2 = A^{11} = \frac{\bar{A}_{11}}{|A|}.$$

Similar distributions exist for the other x's.

This result gives us a simple method of finding the m. g. f. of (x_1-a_1), (x_2-a_2), ..., (x_k-a_k) defined by

$$\phi(\theta_1,\theta_2,..,\theta_k) = E(e^{\sum_{i=1}^{k}\theta_i(x_i-a_i)})$$

(h)
$$= \frac{\sqrt{|A|}}{(2\pi)^{k/2}} \int_{-\infty}^{\infty}\cdots\int e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}(x_i-a_i)(x_j-a_j)+\sum_{i=1}^{k}\theta_i(x_i-a_i)} dx_1\ldots dx_k.$$

Consider the expression

(i)
$$e^{+\frac{1}{2}A_{00}(x_0-a_0)^2}\left(\frac{\sqrt{|A|}\sqrt{\pi}}{\sqrt{A_0}}\right)\left[\frac{\sqrt{A_0}}{(2\pi)^{\frac{k+1}{2}}}\underbrace{\int_{-\infty}\cdots\int}_{k\text{-fold}} e^{-\frac{1}{2}\sum_{i,j=0}^{k} A_{ij}(x_i-a_i)(x_j-a_j)} dx_1\ldots dx_k\right].$$

where $A_{ij} = A_{ji}$ and $A_0 = |A_{ij}|(i,j=0,1,2,..,k)$ and $\| A_{ij}\| (i,j=0,1,2,..,k)$ positive definite.

If we set $A_{0i} = -\theta_i$ and $(x_0-a_0) = 1$, then it will be seen that the expression (i) is exactly the same as that defining $\phi(\theta_1,\theta_2,\ldots,\theta_k)$. But the expression in [ ] is

(j)
$$\frac{\sqrt{B'_{11}}}{\sqrt{2\pi}} e^{-\frac{1}{2}B'_{11}(x_0-a_0)^2}$$

where $B'_{11} = \frac{A_0}{|A|}$. Now by argument presented in §2.94, we may write

$$A_0 = \begin{vmatrix} A_{00} & A_{01} & \cdots & A_{0k} \\ A_{10} & A_{11} & \cdots & A_{1k} \\ \vdots & \vdots & & \\ A_{k0} & A_{k1} & \cdots & A_{kk} \end{vmatrix} = A_{00}|A| - \sum_{i,j=1}^{k} A_{0i}A_{0j}\overline{A}_{ij}$$

Therefore we have

(k)
$$B'_{11} = \frac{A_0}{|A|} = \frac{1}{|A|}\left(A_{00}|A||A| - \sum_{i,j=1}^{k} A_{0i}A_{0j}A^{ij}\right)$$
$$= A_{00} - \sum_{i,j=1}^{k} A_{0i}A_{0j}A^{ij}.$$

Substituting this value of B'_{11} in (j) and the expression for (j) in (i) we find that (i) reduces to

(l)
$$e^{\frac{1}{2}\sum_{i,j=1}^{k} A_{0i}A_{0j}A^{ij}(x_0-a_0)^2}.$$

Setting $x_0-a_0 = 1$, and $A_{0i} = -\theta_i$, we therefore obtain the following result:

Theorem (B): If $x_1, x_2,\ldots,x_k$ are distributed according to the normal multivar-

iate law (b), the m. g. f. of (x_1-a_1), $(x_2-a_2),..,(x_k-a_k)$ is

$$\phi(\theta_1,\theta_2,\ldots,\theta_k) = e^{\frac{1}{2}\sum_{i,j=1}^{k} A^{ij}\theta_i\theta_j}. \tag{m}$$

The argument leading to Theorem (B) may be readily applied to show that any r ($r\leq k$) linearly independent linear functions of (x_i-a_i), $i = 1,2,\ldots$, are distributed according to a normal r-variate distribution. To show this, let

$$L_p = \sum_{i=1}^{k} l_{pi}(x_i-a_i), \qquad p = 1,2,..,r, \tag{n}$$

be the r linearly independent linear functions, i. e., such that there exists no set of constants $C_p(p=1,2,..,r)$ not all zero for which $\sum_{p=1}^{r} l_{pi}C_p = 0$, $i=1,2,..,k$. Let $\phi(\theta_1,\theta_2,\ldots,\theta_r)$ be the m. g. f. of the L_p, i. e.,

$$\phi(\theta_1,\ldots,\theta_r) = \frac{\sqrt{|A|}}{(2\pi)^{\frac{k}{2}}} \int_{-\infty}^{\infty}\cdots\int e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}(x_i-a_i)(x_j-a_j)+\sum_{1}^{k}\theta_p L_p} dx_1\ldots dx_k$$

$$= \frac{\sqrt{|A|}}{(2\pi)^{\frac{k}{2}}} \int_{-\infty}^{\infty}\cdots\int e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}(x_i-a_i)(x_j-a_j)+\sum_{1}^{k} t_i(x_i-a_i)} dx_1\ldots dx_k, \tag{o}$$

where $t_i = \sum_{p=1}^{r}\theta_p l_{pi}$. The value of this integral is given by (1) with $x_o-a_o = 1$, $A_{oi} = -t_i$. Thus

$$\phi(\theta_1,\ldots,\theta_r) = e^{\frac{1}{2}\sum_{i,j=1}^{k} t_i t_j A^{ij}} = e^{\frac{1}{2}\sum_{p,q=1}^{r} B^{pq}\theta_p\theta_q}, \tag{p}$$

where $B^{pq} = \sum_{i,j=1}^{k} A^{ij}l_{pi}l_{qj}$.

Now consider the quadratic form

$$\sum_{p,q=1}^{r} B^{pq}u_p u_q = \sum_{i,j=1}^{r} A^{ij}\Big(\sum_{p} l_{pi}u_p\Big)\Big(\sum_{q} l_{qj}u_q\Big).$$

If $\| A^{ij}\|$ is positive definite and if l_{pi} are linearly independent, then clearly $\| B^{pq}\|$ is positive definite. We therefore have

Theorem (C): Let $x_1,..,x_k$ be distributed according to the normal multivariate law (b), and let $L_p = \sum_{i=1}^{k} l_{pi}(x_i-a_i)$ $(p=1,2,..,r)$ be linearly independent linear functions of the $x_i - a_i$. Then L_p are distributed according to the normal r-variate law

(q) $$\frac{\sqrt{|B|}}{(2\pi)^{\frac{r}{2}}} e^{-\frac{1}{2}\sum_{p,q=1}^{r} B_{pq}L_pL_q}\, dL_1 \ldots dL_r,$$

where $\| B_{pq} \|$ is the inverse of the matrix $\| B^{pq} \|$, and $B^{pq} = \sum_{i,j=1}^{k} A^{ij} l_{pi} l_{qj}$.

Next let us find the conditional p. d. f.

$$f(x_1 | x_2, \ldots, x_k) = \frac{f(x_1, \ldots, x_k)}{g(x_2, \ldots, x_k)},$$

where $g(x_2, \ldots, x_k)$ is the marginal distribution of the last k - 1 variables. Using the marginal distribution found above, where now $\| B^{pq} \| = \| A^{pq} \|$ $(p,q = 2, \ldots, k)$ and also $\| B_{pq} \| = \| A^{(1)}_{pq} \|$, we get

(r) $$f(x_1 | x_2, \ldots, x_k)dx_1 = \frac{\frac{\sqrt{|A|}}{(2\pi)^{k/2}} e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}(x_i - a_i)(x_j - a_j)}\, dx_1 dx_2 \ldots dx_k}{\frac{\sqrt{|B|}}{(2\pi)^{(k-1)/2}} e^{-\frac{1}{2}\sum_{p,q=2}^{k} B_{pq}(x_p - a_p)(x_q - a_q)}\, dx_2, \ldots, dx_k}$$

$$= \frac{\sqrt{A_{11} \cdot |A^{(1)}_{pq}|}\, e^{-\frac{1}{2}|A_{11}[(x_1 - a_1) + \sum_{2}^{k} \frac{A_{1p}}{A_{11}}(x_p - a_p)]^2 + \sum_{2}^{k} A^{(1)}_{pq}(x_p - a_p)(x_q - a_q)|}\, dx_1}{\sqrt{2\pi}\sqrt{|A^{(1)}_{pq}|}\, e^{-\frac{1}{2}\sum_{2}^{k} A^{(1)}_{pq}(x_p - a_p)(x_q - a_q)}}$$

$$= \frac{\sqrt{A_{11}}}{\sqrt{2\pi}} e^{-\frac{1}{2} A_{11}[(x_1 - a_1) + \sum_{2}^{k} \frac{A_{1p}}{A_{11}}(x_p - a_p)]^2}\, dx_1 .$$

Therefore, for fixed values of $x_2, \ldots, x_k$, we have x_1 normally distributed with variance $\frac{1}{A_{11}}$ and mean

(s) $$E(x_1 | x_2, \ldots, x_k) = a_1 - \frac{1}{A_{11}} \cdot \sum_{2}^{k} A_{1p}(x_p - a_p).$$

The regression function for the multivariate normal distribution is linear.

3.3 Pearson System of Distribution Functions

Thus far we have dealt with special distributions which arise under certain specified conditions. Several attempts have been made to develop a general system of distributions which can describe or closely approximate the true distribution of a random variable.

One of these systems derived by Karl Pearson is based upon the differential equation

(a) $$\frac{dy}{dx} = \frac{(x+a)y}{b+cx+dx^2} .$$

Depending on the values given the constants a, b, c, and d we get a wide variety of distribution functions as solutions of the differential equations. We get J-shaped and U-shaped curves, symmetrical and skewed curves, distributions with finite and infinite ranges.

The normal distribution may be obtained as a solution of the differential equation for $c=d=0$ and $b < 0$. This function is Type VII of Pearson's twelve types of solutions.

Another special case we shall be interested in is $d = 0$. Then the equation is

$$\frac{dy}{dx} = \frac{(x+a)y}{b+cx} .$$

Writing this as

$$\frac{dy}{y} = \frac{dx}{c} + \left(\frac{ca-b}{c^2}\right)\frac{dx}{x+\frac{b}{c}} .$$

we see the solution is

$$y = Ke^{\frac{x}{c}}\left(x+\frac{b}{c}\right)^{\frac{ca-b}{c^2}} .$$

Changing the constants, we have

$$y = Ke^{-\beta x}(x+\alpha)^{\nu-1} \qquad \beta>0, \nu>0,$$

where K is chosen so $K\int_{-\alpha}^{\infty} e^{-\beta x}(x+\alpha)^{\nu-1}dx = 1$. This is the Pearson Type III distribution, defined for $-\alpha < x < \infty$.

To determine K we make the indicated integration. Let

$$\frac{z}{\beta} = x + \alpha.$$

Then

$$K\int_{-\alpha}^{\infty} e^{-\beta x}(x+\alpha)^{\nu-1}dx = K'\int_{0}^{\infty} e^{-z}z^{\nu-1}dz,$$

where $K' = Ke^{\beta\alpha}\beta^{-\nu}$. Therefore we choose K' so

$$\frac{1}{K'} = \int_{0}^{\infty} e^{-z}z^{\nu-1}dz.$$

This last integral is an important function of the exponent ν denoted by $\Gamma(\nu)$, the gamma function of ν.

To evaluate $\Gamma(\nu)$ we integrate by parts, using $z^{\nu-1}$ as u and $e^{-z}dz$ as dv.

$$\Gamma(\nu) = -z^{\nu-1}e^{-z}\Big|_{0}^{\infty} + (\nu-1)\int_{0}^{\infty} e^{-z}z^{\nu-2}dz = 0 + (\nu-1)\Gamma(\nu-1).$$

This gives us a recursion or a functional equation for $\Gamma(\nu)$. If ν is an integer,

(b) $$\Gamma(\nu) = (\nu-1)(\nu-2)\ldots2.1\Gamma(1).$$

Since

$$\Gamma(1) = \int_{0}^{\infty} e^{-z}dz = 1,$$

we have for ν an integer,

$$\Gamma(\nu) = (\nu-1)!\ .$$

It is also easy to evaluate $\Gamma(\nu)$ if ν is an integer plus $\frac{1}{2}$. For

$$\Gamma(\tfrac{1}{2}) = \int_{0}^{\infty} z^{-\frac{1}{2}}e^{-z}dz = \int_{-\infty}^{\infty} e^{-t^2}dt = \sqrt{\pi}\ ,$$

and we have

(c) $$\Gamma(\nu) = (\nu-1)(\nu-2)\ldots\tfrac{3}{2}.\tfrac{1}{2}\sqrt{\pi}.$$

In general for $\nu > 0$, $\Gamma(\nu)$ has a finite value, and in any interval (a,b) of values of $\nu (0<a<b)$, $\Gamma(\nu)$ is continuous. $\Gamma(\nu)$ has a minimum for $\nu = 1.46163$.

With this determination of K, the Pearson Type III distribution is

$$\text{(d)} \qquad \frac{e^{-\beta\alpha}\beta^{\nu}}{\Gamma(\nu)}\, e^{-\beta x}(x+\alpha)^{\nu-1}.$$

This distribution for the case $\alpha = 0$ and $\beta = \frac{1}{2}$ is known as the χ^2-distribution with 2ν degrees of freedom and is one of the most important distributions in statistics. It and certain applications will be studied in detail in Chapter V.

It will be convenient at this point to find the moment-generating function of the distribution (d) when $\alpha = 0$. We have

$$\begin{aligned}\phi(\theta) = E(e^{\theta x}) &= \frac{\beta^{\nu}}{\Gamma(\nu)} \int_0^{\infty} e^{\theta x} e^{-\beta x} x^{\nu-1} dx \\ &= \frac{\beta^{\nu}}{\Gamma(\nu)(\beta-\theta)^{\nu}} \int_0^{\infty} e^{-(\beta-\theta)x} [(\beta-\theta)x]^{\nu-1} d[(\beta-\theta)x] \\ &= \frac{\beta^{\nu}}{(\beta-\theta)^{\nu}} .\end{aligned}$$

Therefore, for $\beta - \theta > 0$, we have

$$\phi(\theta) = (1 - \frac{\theta}{\beta})^{-\nu}.$$

For $\beta = \frac{1}{2}$, we have

$$\text{(e)} \qquad \phi(\theta) = (1-2\theta)^{-\nu},$$

which is the m. g. f. for the χ^2-distribution with 2ν degrees of freedom.

Next let us consider the solution of the differential equation (a) when $dx^2 + cx + b$ has two real roots, say g and h ($g<h$), both different from $-a$. Then, using partial fractions we can write the equation as

$$\frac{dy}{dx} = \frac{y(x+a)}{d(x-g)(x-h)} = y(\frac{A}{x-g} - \frac{B}{h-x}) ,$$

where A and B are functions of g, h, a and d which we do not need to determine.

The solution of this equation is

$$y = C(x-g)^{A}(h-x)^{B}, \qquad g \leq x \leq h ,$$

where C is a constant of integration. We wish to determine C so that

$$C\int_g^h (x-g)^A(h-x)^B dx = 1.$$

If we let $x = g + (h-g)v$, the integral becomes

$$C(h-g)^{A+B+1}\int_0^1 v^A(1-v)^B dv.$$

Because we will need the result later, let us evaluate the integral, namely

$$\int_0^1 v^{n_1-1}(1-v)^{n_2-1}dv,$$

which is known as the Beta Function of n_1 and n_2, $B(n_1,n_2)$. We wish to show that this is

$$B(n_1,n_2) = \frac{\Gamma(n_1)\Gamma(n_2)}{\Gamma(n_1+n_2)}. \tag{f}$$

To do this we consider the product $\Gamma(n_1)\,\Gamma(n_2)$, where

$$\Gamma(n_1) = \int_0^\infty x^{n_1-1}e^{-x}dx,$$

and similarly for $\Gamma(n_2)$. Letting $x = s^2$, we get

$$\Gamma(n_1) = 2\int_0^\infty s^{2n_1-1}e^{-s^2}ds.$$

So we can express $\Gamma(n_1)\Gamma(n_2)$ as the double integral

$$\Gamma(n_1)\Gamma(n_2) = 4\int_0^\infty\int_0^\infty s^{2n_1-1}t^{2n_2-1}e^{-s^2-t^2}ds\,dt.$$

If we change to polar coordinates,

$$s = r\cos\theta,$$

$$t = r\sin\theta,$$

this integral over the positive quadrant of the st plane becomes

$$4\int_0^{\frac{\pi}{2}}\int_0^\infty \cos^{2n_1-1}\theta\,\sin^{2n_2-1}\theta\,r^{2n_1+2n_2-1}e^{-r^2}dr\,d\theta.$$

Now

$$2\int_0^{\infty} r^{2n_1+2n_2-1} e^{-r^2} dr = \Gamma(n_1+n_2).$$

If we let $\cos^2\theta = x$, we get

$$2\int_0^{\frac{\pi}{2}} \cos^{2n_1-1}\theta \sin^{2n_2-1}\theta \, d\theta = \int_0^1 x^{n_1-1}(1-x)^{n_2-1} dx = B(n_1,n_2).$$

Combining these results, we have

$$\Gamma(n_1)\Gamma(n_2) = \Gamma(n_1+n_2)B(n_1,n_2),$$

thus proving our desired result.

Therefore, the Type I distribution may be written in the general form

$$(g) \qquad \frac{\Gamma(A+B+2)(x-g)^A(h-x)^B}{\Gamma(A+1)\Gamma(B+1)(h-g)^{A+B+1}} \qquad (g \leq x \leq h).$$

There are twelve types of Pearson distributions. Below are graphed several representative ones.

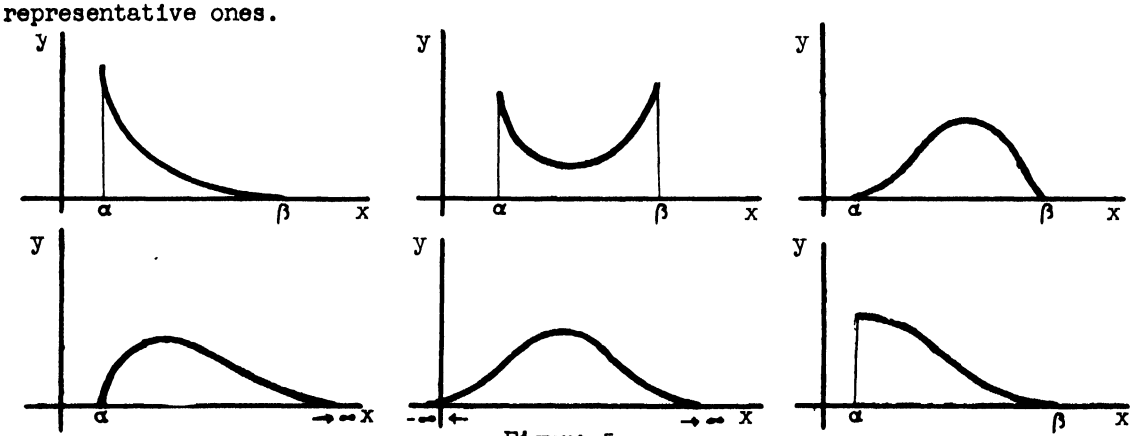

Figure 5

3.4 The Gram-Charlier Series

Another rather general system of distribution functions, known as the Gram-Charlier Series, is based upon the normal distribution and its derivatives. Instead of a number of distributions of different functional forms, this system is composed of an infinite series of terms of a certain kind. Charlier gave a theoretical argument for this system from his development of the hypothesis of elementary errors. We shall regard it, however, as a distribution which has been found satisfactory for fitting or "smoothing" certain empirical distributions.

The <u>generator</u> of this series is the Gaussian or normal distribution. Let

(a) $$\phi_0(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

and let

(b) $$\phi_i(x) = \frac{\partial^i}{\partial x'^i} \phi_0(x), \qquad i = 1,2,\ldots,$$

where $x' = \frac{x-a}{\sigma}$. Then the Gram-Charlier series is

(c) $$f(x) = b_0\phi_0(x) + b_1\phi_1(x) + b_2\phi_2(x) + \ldots = \phi_0(x)\{b_0 - b_1 \frac{x-a}{\sigma} + b_2[\frac{(x-a)^2}{\sigma^2} - 1] - \ldots$$

$$\ldots + (-1)^n b_n H_n(\frac{x-a}{\sigma}) + \ldots\},$$

where $H_n(z)$ is the nth Hermite polynomial

$$H_n(z) = z^n - \frac{n(n-1)}{2} z^{n-2} + \frac{n(n-1)(n-2)(n-3)}{2.4} z^{n-4} - \ldots$$

By choosing the a, σ, and b's properly we obtain a wide variety of distribution functions, which are asymptotic to the x-axis at both ends of the range.

Since

$$\int_{-\infty}^{\infty} f(x)\,dx = b_0,$$

we choose $b_0 = 1$. The mean is

$$\int_{-\infty}^{\infty} xf(x)dx = \sigma b_1 + a.$$

If a in the expression for x' is taken as the mean of the distribution f(x), then $b_1 = 0$. Taking a as the mean of the distribution we find

$$\int_{-\infty}^{\infty} (x-a)^2 f(x)dx = \sigma^2 + 2\sigma^2 b_2.$$

If σ, in the expression for x', is chosen as the standard deviation of f(x) then $b_2 = 0$. It is easily found that the third and fourth moments are

(d) $$\mu_3 = 3!\sigma^3 b_3,$$

(e) $$\mu_4 = \sigma^4(3 - 4!b_4).$$

Similarly, higher moments can be found. Equations (d) and (e) and similar ones for higher moments give equations for determining the b's in terms of moments. The problem of fitting distributions by the use of moments, however, will be discussed in §6.4.

CHAPTER IV

SAMPLING THEORY

4.1 General Remarks

Suppose x is a random variable with c. d. f. F(x). In accordance with the statement made at the end of §2.3, we define a random sample O_n of size n of values of x from a population with c. d. f. F(x) as a set of n random variables $x_1, x_2, \ldots, x_n$ with c. d. f.

(a) $$F(x_1) \cdot F(x_2) \cdots\cdots F(x_n).$$

We note that a random sample consists of statistically independent random variables all having the same c. d. f. It is often convenient to think of x_1 as the value of x in the first "drawing" from the population, x_2 as the value of x in the second "drawing", etc.

In the theory of sampling, we are usually interested in c.d.f.'s of one or more functions of the n random variables comprising the sample. Thus, suppose $g(x_1,x_2,\ldots,x_n)$ is such a sample function (Borel measurable). We are interested in determining the c. d. f. of g, i. e., $Pr[g(x_1,x_2,\ldots,x_n) \leq g]$, the value of which is obtained by performing the Stieltjes integration

(b) $$\int_R \cdots \int dF(x_1) \cdots\cdots dF(x_n),$$

where R is the region in the n-dimensional space of the x's for which $g(x_1,x_2,\ldots,x_n) \leq g$.

Similarly, if $g_i(x_1,x_2,\ldots,x_n)$ $(i = 1,2,\ldots,k)$, $k \leq n$, are k Borel measurable functions, we are interested in determining $Pr(g_i(x_1,x_2,\ldots,x_n) \leq g_i \ (i = 1,2,\ldots,k))$.

The random variable x may be a vector with r components, say $x^{(1)},x^{(2)},\ldots,x^{(r)}$, with c. d. f. $F(x^{(1)},x^{(2)},\ldots,x^{(r)})$. In this case the sample O_n would consist of n random vectors $(x_\alpha^{(1)},x_\alpha^{(2)},\ldots,x_\alpha^{(r)})$, $\alpha = 1,2,\ldots,n$, (a total of nk vandom variables) with c. d. f

$$\prod_{\alpha=1}^{n} F(x_\alpha^{(1)},x_\alpha^{(2)},\ldots,x_\alpha^{(r)}).$$

Again, the sampling problem is to determine the c. d. f. of one or more (Borel measurable)

functions of the nk random variables involved. For example, here one may wish to determine the probability theory of such functions as $\bar{x}^{(i)}$, $\sum_{\alpha=1}^{n} x_\alpha^{(i)} x_\alpha^{(j)}$, $\sum_{\alpha=1}^{n} (x_\alpha^{(i)} - \bar{x}^{(i)}) \cdot (x_\alpha^{(j)} - \bar{x}^{(j)})$, where $\bar{x}^{(i)} = \frac{1}{n} \sum_{\alpha=1}^{n} x_\alpha^{(i)}$, i, j = 1,2,...,r, and other symmetrical functions.

In mathematical statistics one is usually interested in relatively simple sample functions, such as averages, ratios, sums of squares, correlation coefficients, etc. One is able to obtain simple expressions for sampling distributions for such functions only in certain special cases which will be considered in this and in later chapters. However, one is able to obtain moments of some of the simpler g functions such as averages, average sum of squares, etc., under broader conditions. Some of these cases will also be considered.

4.2 Application of Theorems on Mean Values to Sampling Theory

This section consists of the application of results of §§2.71-2.75 to cases of interest in sampling theory. No assumptions are made about the population distribution except the existence of first and second moments.

4.21 Distribution of Sample Mean

Let $O_n:(x_1,x_2,\ldots,x_n)$ be a sample from a population with an arbitrary distribution for which the first moment $\mu_1' = a$ exists. Let $\bar{x}$ be the mean of the sample,

$$\bar{x} = \sum_{i=1}^{n} x_i/n.$$

Then from equation (b) of §2.74, we have that the expected value of $\bar{x}$ is

$$E(\bar{x}) = \sum_{i=1}^{n} a_i/n = a,$$

since $a_i = E(x_i) = a$. If furthermore the population distribution F(x) has a finite variance σ^2, then since each x_i has the c. d. f. $F(x_i)$, and the x_i are mutually independent, we get from (d) of §2.74 that the variance of $\bar{x}$ is

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{n} \sigma^2/n^2 = \sigma^2/n.$$

We gather these results into

Theorem (A): If $\bar{x}$ is the mean of a sample of size n from a population with arbitrary c. d. f. F(x), then if the mean a of F(x) exists,

$$E(\bar{x}) = a,$$

and if F(x) has finite variance σ^2, the variance of $\bar{x}$ is

$$\sigma_{\bar{x}}^2 = \sigma^2/n.$$

Having computed the mean and variance of $\bar{x}$ we may now apply Tchebycheff's inequality (§2.71):

(a) $$Pr(|\bar{x}-a| > \delta\sigma/n) \leq 1/\delta^2.$$

Let ϵ be an arbitrary positive number, and define δ from $\delta\sigma/n = \epsilon$. Then (a) may be written

(b) $$Pr(|\bar{x}-a| > \epsilon) \leq \sigma^2/n^2\epsilon^2.$$

Now a random variable X_n which is defined for $n = 1,2,3,\ldots$, is said to converge stochastically to a value A if

$$Pr(|X_n-A|<\epsilon)\longrightarrow 0 \text{ as } n\longrightarrow\infty \text{ for every fixed } \epsilon > 0.$$

Letting $n \longrightarrow \infty$ in (b) we get

Theorem (B): For an arbitrary population with finite variance, the sample mean converges stochastically to the population mean.

For the sample of size n let $G_n(\bar{x})$ be the c. d. f. of $\bar{x}$. From theorem (A) we see that the limiting form of G_n is the step function

$$\lim_{n\longrightarrow\infty} G_n(\bar{x}) = \begin{cases} 0 \text{ for } \bar{x} < a, \\ 1 \text{ for } \bar{x} > a. \end{cases}$$

In order to "spread out" again the probability which all "piles up" at $\bar{x} = a$, we might consider the distribution of $z = (\bar{x}-a)/h(n)$, where the function h(n) is chosen so as to keep the variance of z from approaching zero. From (d) of §2.74, we see that

$$\sigma_z^2 = \sigma_{\bar{x}}^2/[h(n)]^2 = \sigma^2/n[h(n)]^2.$$

Hence if we choose $h(n) = n^{-\frac{1}{2}}\sigma$, the variable z has zero mean and unit variance for all n. A beautiful result about the limiting distribution of z as $n \longrightarrow \infty$, regardless of the population distribution, is contained in the central limit theorem:

Theorem (C): For an arbitrary population with mean a and finite variance σ^2, the c. d. f. $G_n(z)$ of

$$z = (\bar{x}-a)\sqrt{n}/\sigma$$

approaches the normal distribution $N(0,1)$ as $n \longrightarrow \infty$,

(c) $$G_n(z) \longrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2}\, dt \quad \text{as } n \longrightarrow \infty,$$

uniformly in z.

We make the proof for the case where the m. g. f. $\psi(\theta)$ of the original distribution exists for $|\theta|<h$, $h>0$. Then for $|\theta|<h$, the m. g. f. $\phi(\theta)$ of $y = (x-a)/\sigma$ also exists, for $\phi(\theta) = e^{-a\theta/\sigma}\psi(\theta/\sigma)$. Finally, let $\Phi(\theta)$ be the m. g. f. of z:

$$\Phi(\theta) = E(e^{\theta z}) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp[\theta \sum_{1}^{n} (x_1-a)/\sqrt{n}\sigma]dF(x_1)dF(x_2)\cdots dF(x_n)$$

$$= \left[\int_{-\infty}^{+\infty} \exp\,[\theta(x-a)/\sqrt{n}\sigma]dF(x)\right]^n = |\phi(\theta/\sqrt{n})|^n.$$

Now

$$\phi(u) = \phi(0) + u\phi'(0) + \tfrac{1}{2}u^2\phi''(u_1),$$

where $0 < u_1 < u < h$ if $u > 0$, and $-h < u < u_1 < 0$ if $u < 0$. $\phi''(u)$ is continuous at $u = 0$, hence $\phi''(u) = \phi''(0) + \eta(u)$, where $\eta(u) \longrightarrow 0$ as $u \longrightarrow 0$. We recall $\phi^{(i)}(0)$ is the i^{th} moment if y about the origin, so $\phi(0) = 1$, $\phi'(0) = 0$, $\phi''(0) = 1$, and

(d) $$\Phi(\theta) = \left[1 + \frac{\theta^2}{2n}\,[1+\eta(\frac{\theta_1}{\sqrt{n}})]\right]^n,$$

where $0 < \theta_1 < \theta < h\sqrt{n}$ or $-h\sqrt{n} < \theta < \theta_1 < 0$. Now choose any θ and hold it fixed. (d) is valid for $n > \theta^2/h^2$. Letting $n \longrightarrow \infty$, for every fixed θ,

$$\lim_{n\longrightarrow\infty} \Phi(\theta) = e^{\frac{1}{2}\theta^2},$$

which is the m. g. f. for $N(0,1)$. Therefore from Theorem (C) of §2.91, the limiting distribution of $G_n(z)$ is given by (c) above.

While the above proof based on the generating function can be shortened, we have purposely given it in a way which permits of generalization to distributions of which it is assumed only that the second moment exists. In this general case one employs instead of the m. g. f. the characteristic function $\Psi(t)$ of the distribution, which is related to the generating function $\psi(\theta)$ by $\Psi(t) = \psi(it)$. This always exists for all real t. The argument follows the above step by step and at the end one appeals to a theorem analogous to (C) of §2.91, which states that if the limit of the characteristic function is the charac-

teristic function of some continuous c. d. f. $F^*(x)$ then the limit of the c. d. f. is $F^*(x)$ uniformly for all x.

4.22 Expected Value of Sample Variance

For the sample $O_n:(x_1,x_2,\ldots,x_n)$, call S the sum of squared deviations from the sample mean,

$$S = \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2-n\bar{x}^2.$$

Recalling that E is a linear operator, we get

$$E(S) = \sum_{i=1}^{n}E(x_i^2) - nE(\bar{x}^2).$$

Now if the population distribution F(x) has mean a and finite variance σ^2,

$$E(x_i^2) = [\mu_2' \text{ of } F(x)] = \sigma^2 + a^2,$$

$$E(\bar{x}^2) = [\mu_2' \text{ of c.d.f. of } \bar{x}] = \sigma_{\bar{x}}^2 + a^2 = a^2 + \sigma^2/n.$$

Thus

$$E(S) = (n-1)\sigma^2.$$

We note that $E(S/n) \neq \sigma^2$, but if we define

$$s^2 = S/(n-1),$$

then

$$E(s^2) = \sigma^2.$$

4.3 Sampling from a Finite Population

Suppose that a population has a finite number N of elements, each characterized by a number $x = x^{(i)}$, $i = 1,2,\ldots,N$, and that we draw a random sample $O_n:(x_1,x_2,\ldots,x_n)$ without replacement. The sample may be represented by a point $(x_1,x_2,\ldots,x_n)$ in n dimensions, the possible values of x_α being $x^{(1)}, x^{(2)},\ldots,x^{(N)}$, $\alpha = 1,2,\ldots,n$. To simplify the discussion, let us assume that the values of the $x^{(i)}$ are distinct, $i = 1,2,\ldots,N$. Then $\Pr(x_\alpha=x_\beta \text{ for } \alpha\neq\beta) = 0$. Hence we may think of the range of the sample point being all points of the lattice $x_\alpha = x^{(1)},x^{(2)},\ldots,x^{(N)}$, $\alpha = 1,2,\ldots,n$, but we must ascribe to any point for which $x_\alpha = x_\beta$, $\alpha\neq\beta$, the probability zero. By a random sample we mean that all points of this lattice, barring the exceptional points just mentioned, have the same probability p. To enumerate the points with probability p, we note that to obtain such a point, we may choose x_1 in N ways, x_2 in N-1 ways, ..., x_n in N-n+1 ways. The number of

points with probability p is thus N(N-1)...(N-n+1). Since the total probability of the points of the lattice must add up to unity, we have

(a) $$p = [N(N-1)\ldots(N-n+1)]^{-1},$$

$$p(x_1,x_2,\ldots,x_n) = p\,\delta_{x_1x_2\ldots x_n},$$

where

$$\delta_{x_1x_2\ldots x_n} \begin{cases} 0 \text{ if any two } x_\alpha \text{ are equal,} \\ 1 \text{ if all } x_\alpha \text{ are distinct.} \end{cases}$$

Define the mean $\underline{a}$ and the variance σ^2 of the population from

$$a = \sum_{i=1}^{N} x^{(i)}/N, \qquad \sigma^2 = \sum_{i=1}^{N}(x^{(i)}-a)^2/N.$$

Here, we shall consider the problem of determining the mean and variance of the mean of a random sample from this population. Let $\bar{x}$ be the sample mean,

$$\bar{x} = \sum_{\alpha=1}^{n} x_\alpha/n.$$

We note that the x_α are <u>not</u> independent, --it will later be seen that the correlation between x_α and x_β is not zero, --but we may nevertheless use the formula (f) of §2.74, as pointed out there. Thus

(b) $$E(\bar{x}) = \sum_{\alpha=1}^{n} E(x_\alpha)/n.$$

To calculate $E(x_\alpha)$ we desire the marginal distribution of x_α. Suppose $\alpha = 1$. Then $Pr(x_1=x^{(i)})$ is the sum of the probability over all lattice points for which $x_1 = x^{(i)}$, that is, it is p times the number of lattice points for which $x_1 = x^{(i)}$, and no two of $x_1,x_2,\ldots,x_n$ are equal. To compute this number note that we may choose x_1 in only one way, $x_1 = x^{(i)}$, then x_2 in N-1 ways, $x_2 \neq x^{(i)}$, then x_3 in N-2 ways, $x_3 \neq x^{(i)}$ or x_2, etc.; so the desired number is (N-1)(N-2)...(N-n+1). The marginal probability of x_1 is thus seen to be

$$Pr(x_1=x^{(i)}) = p(N-1)(N-2)\ldots(N-n+1) = 1/N$$

from (a). We get

$$E(x_1) = \sum_{i=1}^{N} x^{(i)} Pr(x_1=x^{(i)}) = \sum_{i=1}^{N} x^{(i)}/N = a.$$

Similarly,

$$E(x_\alpha) = a, \qquad \alpha = 1,2,\ldots,n,$$

and substituting in (b), we find

$$E(\bar{x}) = a.$$

To calculate $\sigma_{\bar{x}}^2$ we use formula (c) of §2.74,

(c) $$\sigma_{\bar{x}}^2 = \sum_{\alpha,\beta=1}^{n} \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta/n^2.$$

Employing again the marginal distribution of x_α, we get for the variance of x_α

$$\sigma_\alpha^2 = E(x_\alpha^2) - [E(x_\alpha)]^2 = \sum_{i=1}^{N} [x^{(i)}]^2 \Pr(x_\alpha = x^{(i)}) - a^2 = \sum_{i=1}^{N} [x^{(i)}]^2/N - a^2,$$

(d) $$\sigma_\alpha = \sigma.$$

To find $\rho_{\alpha\beta}$ for $\alpha \neq \beta$, we use the joint marginal distribution of x_α and x_β. To simplify the notation, let $\alpha = 1$, $\beta = 2$. Then $\Pr(x_1 = x^{(i)}, x_2 = x^{(j)}; i \neq j)$ is p times the number of points for which $x_1 = x^{(i)}$, $x_2 = x^{(j)} \neq x^{(i)}$, and no two of $x_1, x_2, \ldots, x_n$ are equal. To enumerate these points, note that we may choose x_1, x_2 in only one way, then x_3 in N-2 ways, x_4 in N-3 ways, etc. Hence

$$\Pr(x_1 = x^{(i)},\ x_2 = x^{(j)}, i \neq j) = p(N-2)(N-3)\ldots(N-n+1) = [N(N-1)]^{-1},$$

$$\sigma_1\sigma_2\rho_{12} = E[(x_1-a)(x_2-a)] = \sum_{\substack{i,\,j \\ i\neq j}} (x^{(i)}-a)(x^{(j)}-a)\Pr(x_1 = x^{(i)}, x_2 = x^{(j)})$$

$$= \sum_i (x^{(i)}-a) \sum_{\substack{j \\ j\neq i}} (x^{(j)}-a)/[N(N-1)]$$

$$= \sum_i (x^{(i)}-a)[-x^{(i)} + \sum_j (x^{(j)}-a)]/[N(N-1)]$$

$$= \sum_i (x^{(i)}-a)(-x^{(i)})/[N(N-1)]$$

$$= -\sum_i [(x^{(i)})^2/N - ax^{(i)}/N]/(N-1) = -\sigma^2/(N-1),$$

$$\rho_{12} = -1/(N-1).$$

Likewise,

(e) $$\rho_{\alpha\beta} = -1/(N-1) \text{ if } \alpha \neq \beta.$$

Combining (c), (d), (e), we have

$$n^2\sigma_{\bar{x}}^2 = \sum_{\alpha=1}^{n}\sigma_\alpha^2 + 2\sum_{\substack{\alpha,\beta \\ \alpha<\beta}}\rho_{\alpha\beta}\sigma_\alpha\sigma_\beta$$

$$= n\sigma^2 + 2(1+2+3+\ldots+n-1)[-1/(N-1)]\sigma^2$$

$$= n\sigma^2 - n(n-1)\sigma^2/(N-1),$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

We note that for $n = N$, $\sigma_{\bar{x}} = 0$, that $\sigma_{\bar{x}}$ is a monotonic increasing function of N, and that as $N \longrightarrow \infty$, $\sigma_{\bar{x}}^2 \longrightarrow \sigma^2/n$ for fixed n.

4.4 Representative Sampling

Suppose we have a population π consisting of k mutually exclusive sub-populations π_i, each with c. d. f. $F_i(x)$, that is,

$$F_i(x) = \Pr(X \leq x \mid X \text{ from } \pi_i).$$

If X is drawn at random from π, let

$$p_i = \Pr(X \text{ from } \pi_i), \qquad \sum_{i=1}^{k} p_i = 1.$$

To find the c. d. f. of X we may proceed as follows:

$$F(x) = \Pr(X \leq x) = \sum_{i=1}^{k}\Pr(X \text{ from } \pi_i)\cdot\Pr(X \leq x \mid X \text{ from } \pi_i) = \sum_{i=1}^{k} p_i F_i(x).$$

Denoting the mean of F(x) by $\underline{a}$, and its variance by σ^2, we calculate

$$dF(x) = \sum_{i=1}^{k} p_i\, dF_i(x),$$

$$a = \int_{-\infty}^{+\infty} x\, dF(x) = \sum_{i=1}^{k} p_i \int_{-\infty}^{+\infty} x\, dF_i(x) = \sum_{i=1}^{k} p_i a_i,$$

where a_i is the mean of $F_i(x)$.

$$\sigma^2 + a^2 = \int_{-\infty}^{+\infty} x^2 dF(x) = \sum_{i=1}^{k} p_i \int_{-\infty}^{+\infty} x^2 dF_i(x) = \sum_{i=1}^{k} p_i(\sigma_i^2 + a_i^2),$$

where σ_i^2 is the variance of $F_i(x)$. This may be written

$$\sigma^2 = \sum_{i=1}^{k} p_i[\sigma_i^2 + (a_i - a)^2].$$

From §4.21 we have that if $\bar{x}$ is the mean of a sample of size n drawn at random from π, then

$$E(\bar{x}) = a,$$

(a)
$$\sigma_{\bar{x}}^2 = \sigma^2/n = \sum_{i=1}^{k} p_i[\sigma_i^2 + (a_i - a)^2]/n.$$

4.41 Sampling when the p_i are known

We suppose the probabilities p_i are known (the means a_i are assumed throughout to be unknown). Let us draw a sample O_n consisting of the following sub-samples: $O^{(1)}$ (n_1 elements from π_1), $O^{(2)}$ (n_2 elements from π_2), .., $O^{(k)}$ (n_k elements from π_k); $\sum_{i=1}^{k} n_i = n$. Call $\bar{x}_R$ the mean of O_n, and $\bar{x}_i$ the mean of $O^{(i)}$. Then

(a)
$$\bar{x}_R = \sum_{i=1}^{k} \bar{x}_i n_i/n,$$

$$E(\bar{x}_R) = \sum_{i=1}^{k} E(\bar{x}_i) n_i/n = \sum_{i=1}^{k} a_i n_i/n.$$

If we use $\bar{x}_R$ as an estimate of the mean a of π, we would like to have

$$E(\bar{x}_R) = a = \sum_{i=1}^{k} a_i p_i.$$

Since we do not know the a_i, we require

$$\sum_{i=1}^{k} a_i p_i = \sum_{i=1}^{k} a_i n_i/n$$

for all a_i, and this uniquely determines the n_i as np_i.

If $n_i = np_i$, then O_n is called a representative sample from π. The advantages of representative sampling over random sampling from π are implicit in

Theorem (A): The variance $\sigma_{\bar{x}_R}^2$ of the mean $\bar{x}_R$ of a representative sample and the variance $\sigma_{\bar{x}}^2$ of the mean $\bar{x}$ of a random sample of the same size have the following relationship:

$$\sigma^2_{\bar{x}_R} \leq \sigma^2_{\bar{x}},$$

the equality holding only when all a_i are equal.

To prove the theorem, we calculate

(b) $$\sigma^2_{\bar{x}_R} = \sum_{i=1}^{k} \sigma^2_{\bar{x}_i} (n_i/n)^2$$

from (a) and the mutual independence of the $\bar{x}_i$. Now

$$\sigma^2_{\bar{x}_i} = \sigma_i^2/n_i = \sigma_i^2/np_i.$$

Therefore

$$\sigma^2_{\bar{x}_R} = \sum_{i=1}^{k} (\sigma_i^2/np_i)p_i^2 = \sum_{i=1}^{k} \sigma_i^2 p_i/n.$$

Hence (a) of §4.4 may be written

$$\sigma^2_{\bar{x}} = \sigma^2_{\bar{x}_R} + \sum_{i=1}^{k} p_i(a_i-a)^2/n,$$

and the theorem follows.

4.42 Sampling when the σ_i are also known

We employ the same notation as in §4.41. If we use the mean $\bar{x}_R$ of the sample to estimate $\underline{a}$, we have just seen that the n_i are uniquely determined by the requirement

$$E(\bar{x}_R) = a.$$

Suppose however that we use as an estimate of $\underline{a}$ the statistic

(a) $$y = \sum_{i=1}^{k} c_i \bar{x}_i.$$

How should we choose the n_i, for fixed $n = \sum_{i=1}^{k} n_i$, so that

(b) $$E(y) = a,$$

and σ^2_y is minimum (for the class of statistics satisfying (a) and (b))? The method of §4.41 shows that we must take $c_i = p_i$. Then

(c) $$\sigma^2_y = \sum_{i=1}^{k} p_i^2 \sigma^2_{\bar{x}_i} = \sum_{i=1}^{k} p_i^2 \sigma_i^2/n_i.$$

The problem is now to find the n_i which minimize (c) subject to the condition that $\sum_{i=1}^{k} n_i = n$. Treating the n_i as though they were continuous variables, and following the method of Lagrange, we form

$$g(n_1,n_2,\ldots,n_k;\lambda) = \sum_{i=1}^{k} p_i^2\sigma_i^2/n_i + \lambda\left(\sum_{i=1}^{k} n_i - n\right),$$

and set

$$\partial g/\partial n_i = 0, \qquad i = 1,2,\ldots,k.$$

We get

$$-p_i^2\sigma_i^2/n_i^2 + \lambda = 0,$$

$$n_i = p_i\sigma_i/\lambda^{\frac{1}{2}}. \tag{d}$$

To evaluate $\lambda^{\frac{1}{2}}$ sum the equations (d) for $i = 1,2,\ldots,k$, and solve for $\lambda^{\frac{1}{2}}$:

$$\lambda^{\frac{1}{2}} = \sum_{j=1} p_j\sigma_j/n.$$

The minimizing n_i are thus

$$n_i = np_i\sigma_i/\left(\sum_{j=1}^{k} p_j\sigma_j\right).$$

Putting these back in (c), we find the minimum variance to be

$$\sigma_y^2 = \left(\sum_{i=1}^{k} p_i\sigma_i\right)^2/n.$$

With the help of the Schwartz inequality,

$$\left(\sum_{i=1}^{k} a_ib_i\right)^2 \leq \left(\sum_{i=1}^{k} a_i\right)\left(\sum_{i=1}^{k} b_i\right),$$

(the equality holding only if the a_i are proportional to the b_i), where we let $a_i = p_i^{\frac{1}{2}}$, $b_i = p_i^{1/2}\sigma_i$, we obtain

Theorem (A): $\sigma_y^2 \leq \sigma_{\bar{x}_R}^2$,

the equality holding only if all σ_i are equal.

4.5 Sampling Theory of Order Statistics

4.51 Simultaneous Distribution of any k Order Statistics. Suppose $O_n:(x_1,x_2,\ldots,x_n)$ is a sample of size n from a population with probability element $f(x)dx$, and that $x_1,x_2,\ldots,x_n$ are arranged in ascending order of magnitude. These ordered values of x will be referred to as order statistics, more specifically, x_α will be called the α^{th} order

statistic. Let $r_1, r_2, \ldots, r_k$ be k integers such that $1 \leq r_1 < r_2 < \ldots < r_k \leq n$. The problem to be considered here is that of finding the probability element of $x_{r_1}, x_{r_2}, \ldots, x_{r_k}$, i. e.

(a) $$f(x_{r_1}, x_{r_2}, \ldots, x_{r_k})\, dx_{r_1} dx_{r_2} \ldots dx_{r_k}.$$

Let $I_1, I_2, I_3, \ldots, I_{2k+1}$ be the 2k+1 intervals

(b) $$(-\infty, x_{r_1}), (x_{r_1}, x_{r_1}+dx_{r_1}), (x_{r_1}+dx_{r_1}, x_{r_2}), (x_{r_2}, x_{r_2}+dx_{r_2}), \ldots, (x_{r_k}+dx_{r_k}, +\infty),$$

and let

$$\int_{I_i} f(x)dx = q_i, \qquad i = 1,2,..,2k+1. \quad (\sum_1^{2k+1} q_i = 1.)$$

The problem of finding the probability element (a) is identical with that of finding the probability (to terms of order $dx_{r_1} dx_{r_2} \ldots dx_{r_k}$) that if a sample of n elements is drawn from a multinomial population with classes $I_1, I_2, \ldots, I_{2k+1}$ then r_1-1 elements will fall in I_1, 1 element in I_2, r_2-r_1-1 elements in I_3, 1 element in I_4, ..., $n - r_k-1$ elements in I_{2k+1}. It follows from the multinomial law (§3.12) that the probability of such a partition is

(c) $$\frac{n!}{(r_1-1)!1!(r_2-r_1-1)!1!\ldots(n-r_k-1)!}\, q_1^{r_1-1} q_2^{1} q_3^{r_2-r_1-1} q_4^{1} \ldots q_{2k+1}^{n-r_k-1}.$$

Substituting the values of the q_i, and noting that, to within terms of order dx_{r_i},

(d) $$\int_{x_{r_i}}^{x_{r_i}+dx_{r_i}} f(x)dx = f(x_{r_i})dx_{r_i} \quad \text{and} \quad \int_{x_{r_i}+dx_{r_i}}^{x_{r_{i+1}}} f(x)dx = \int_{x_{r_i}}^{x_{r_{i+1}}} f(x)dx,$$

we have

(e) $$f(x_{r_1}, x_{r_2}, \ldots, x_{r_k})dx_{r_1} dx_{r_2} \ldots dx_{r_k}$$

$$= \frac{n!}{(r_1-1)!(r_2-r_1-1)!\ldots(n-r_k-1)!} \left(\int_{-\infty}^{x_{r_1}} f(x)dx\right)^{r_1-1} \left(\int_{x_{r_1}}^{x_{r_2}} f(x)dx\right)^{r_2-r_1-1} \ldots \left(\int_{x_{r_k}}^{\infty} f(x)dx\right)^{n-r_k-1}$$

$$f(x_{r_1})dx_{r_1} \ldots f(x_{r_k})dx_{r_k}.$$

The distribution function (e) has many applications, some of which will now be considered briefly.

4.52 Distribution of Largest (or Smallest) Variate

In this case $k = 1$, $r_k = n$; (e) of §4.51 then becomes the probability element of the largest element x_n,

$$n\Big(\int_{-\infty}^{x_n} f(x)dx\Big)^{n-1} f(x_n)dx_n ,$$

a similar expression holding for the probability element of the smallest element.

4.53 Distribution of Median

In this case let the number of elements in the sample be $2n + 1$. We would then have $k = 1$, $r_k = n + 1$, and (e) of §4.51 will be the probability element of the sample median x_{n+1}. Denoting the median by $\tilde{x}$, we have

$$\text{(a)} \qquad \frac{(2n+1)!}{(n!)^2}\Big(\int_{-\infty}^{\tilde{x}} f(x)dx\Big)^n\Big(\int_{\tilde{x}}^{\infty} f(x)dx\Big)^n f(\tilde{x})d\tilde{x}.$$

The asymptotic distribution of the median for large n may be derived from (a). If $\tilde{x}_o$ is the population median then $\int_{-\infty}^{\tilde{x}_o} f(x)dx = \frac{1}{2}$. Therefore

$$\int_{-\infty}^{\tilde{x}} f(x)dx = \frac{1}{2} + \int_{\tilde{x}_o}^{\tilde{x}} f(x)dx \text{ and } \int_{\tilde{x}}^{\infty} f(x)dx = \frac{1}{2} - \int_{\tilde{x}_o}^{\tilde{x}} f(x)dx,$$

and hence (a) may be written as

$$\text{(b)} \qquad \frac{(2n+1)!}{2^{2n}(n!)^2}\Big[1-4\Big(\int_{\tilde{x}_o}^{\tilde{x}} f(x)dx\Big)^2\Big]^n f(\tilde{x})d\tilde{x}.$$

We may write $\int_{\tilde{x}_o}^{\tilde{x}} f(x)dx = \bar{f}(\tilde{x}-\tilde{x}_o)$, where

$$\min_{x\in I} f(x) \leq \bar{f} \leq \max_{x\in I} f(x),$$

and I is the interval $(\tilde{x}_o,\tilde{x})$ or $(\tilde{x},\tilde{x}_o)$.

Let $\sqrt{n}(\tilde{x}-\tilde{x}_o) = y$. Then (b) becomes

$$\text{(c)} \qquad \frac{(2n+1)!}{2^{2n}(n!)^2}\Big(1 - \frac{4f^{-2}y^2}{n}\Big)^n f\Big(\frac{y}{\sqrt{n}} + \tilde{x}_o\Big)dy/\sqrt{n}.$$

We now choose any value of y, hold it fixed, and let $n \to \infty$. If $f(x)$ is continuous at $x = \tilde{x}_0$ and $f(\tilde{x}_0) \neq 0$, then $f(\tilde{x}_0+y/\sqrt{n}) \to f(\tilde{x}_0)$, $\overline{f} \to f(\tilde{x}_0)$, and with the help of Stirling's formula for the factorials, we thus get as the limit of (c) as $n \to \infty$,

$$\text{(d)} \qquad \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}y^2/\sigma_y^2} dy,$$

where $\sigma_y^2 = 1/8[f(x_0)]^2$. Hence the median $\tilde{x}$ in samples of size $2n + 1$ is asymptotically normally distributed with mean $\tilde{x}_0$ and variance $1/8n[f(\tilde{x}_0)]^2$. It is of interest to note that this asymptotic distribution depends only on the $\tilde{x}_0$ and $f(\tilde{x}_0)$ of the population.

Example: For the normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-a)^2}$$

we have $\tilde{x}_0 = a$, $f(\tilde{x}_0) = 1/\sqrt{2\pi}\sigma$. Therefore, the variance $\sigma_{\tilde{x}}^2$ of $\tilde{x}$ in samples of size $2n + 1$ from a normal distribution with variance σ^2 is $\pi\sigma^2/4n$, approximately. It will be recalled from §4.21 that the variance $\sigma_{\overline{x}}^2$ of the mean of a sample of size $2n + 1$ is $\sigma^2/(2n+1)$. Hence, for large samples from a normal population, the mean has smaller variance than the median.

In a similar manner one could treat the problem of finding the sampling distribution of the lower quartile of a sample (the $(n+1)$st element in rank order in a sample of size $4n + 3$), and other particular order statistics.

4.54 Distribution of Sample Range

The joint distribution of the largest and smallest values of x in the sample is given by (e) of §4.51 with $k = 2$, $r_1 = 1$, $r_2 = n$. We have

$$\text{(a)} \qquad n(n-1)\Big(\int_{x_1}^{x_n} f(x)dx\Big)^{n-2} f(x_1)f(x_n)dx_1 dx_n.$$

To obtain the distribution of the sample range R, we make the following transformation

$$\text{(b)} \qquad \begin{aligned} x_n - x_1 &= R \\ x_1 &= S \end{aligned}$$

and integrate the resulting distribution with respect to S.

Example: Suppose x has the rectangular distribution

$$\text{(c)} \qquad \begin{aligned} f(x) &= 1/r, && 0 < x < r, \\ &= 0, && \text{otherwise.} \end{aligned}$$

We have for (a),

$$\text{(d)} \qquad n(n-1)r^{-n}(x_n-x_1)^{n-2}dx_1 dx_n.$$

Applying transformation (b) and integrating with respect to S from 0 to r - R, we obtain as the probability element of the range in samples of size n from the rectangular distribution

(e) $$n(n-1)r^{-n}R^{n-2}(r-R)dR.$$

4.55 Tolerance Limits

The joint distribution of the smallest and largest values of x in the sample is given by (a) of §4.54. Now suppose we set

(a) $$\int_{-\infty}^{x_1} f(x)dx = u, \qquad \int_{x_1}^{x_n} f(x)dx = v.$$

We have

$$\frac{\partial(x_1,x_n)}{\partial(u,v)} = \left[\frac{\partial(u,v)}{\partial(x_1,x_n)}\right]^{-1} = \frac{1}{f(x_1)\cdot f(x_n)},$$

and hence the joint distribution of u and v is

(b) $$n(n-1)\, v^{n-2}dudv,$$

and the region of non-zero probability density is the triangle bounded by u=0, v=0, u+v=1. The probability element (b) clearly does not depend on the probability density function f(x). Integrating with respect to u from 0 to 1-v, we find the probability element of v to be

(c) $$n(n-1)\, v^{n-2}(1-v)dv.$$

It will be seen that v is the amount of the probability in the distribution f(x) included between x_1 and x_n (or statistically speaking, it is the proportion of the population included between x_1 and x_n, i. e. between the least and greatest values of a sample of size n). From expression (c) one can determine the sample size n such that the probability is ϵ that at least $100\beta\%$ of the population will be included between the least and greatest value of the sample. Such a value of n would be obtained by solving the following equation for n:

(d) $$n(n-1)\int_{\beta}^{1} v^{n-2}(1-v)dv = \epsilon,$$

or

(e) $$n\beta^{n-1} - (n-1)\beta^{n} = 1 - \epsilon.$$

Example: For $\epsilon = .95$ and $\beta = .99$, we find n = 130. Thus, if a sample of 130 cases is drawn from a population in which the random variable x is continuous, the probability is .95 that the least and greatest values of x in the sample will include

at least 99% of the population.

x_1 and x_n are examples of tolerance limits. More generally, two functions of the sample values, say $L_1(x_1,\ldots,x_n)$ and $L_2(x_1,x_2,\ldots,x_n)$, will be called 100β% distribution-free tolerance limits at probability level ϵ if

$$(f) \qquad \Pr\left(\int_{L_1}^{L_2} f(x)dx \geq \beta\right) = \epsilon,$$

for all possible probability density functions $f(x)$.

If the functional form of $f(x)$ is known but depends on one or more parameters $\theta_1,\theta_2,\ldots,\theta_h$ and if L_1 and L_2 are such that (f) holds for all possible values of the parameters we shall call L_1 and L_2 100β% parameter-free tolerance limits at probability level ϵ.

If we denote by $u_1, u_2,\ldots,u_k$ the quantities

$$\int_{-\infty}^{x_{r_1}} f(x)dx, \int_{x_{r_1}}^{x_{r_2}} f(x)dx, \ldots, \int_{x_{r_{k-1}}}^{x_{r_k}} f(x)dx,$$

respectively, it is easy to verify in a manner similar to our treatment of the distribution of u and v, that the probability element of $u_1,u_2,\ldots,u_k$ is

$$(g) \qquad \frac{n!}{(r_1-1)!(r_2-r_1-1)!\ldots(n-r_k-1)!}\, u_1^{r_1-1} u_2^{r_2-r_1-1}\ldots u_k^{r_k-r_{k-1}-1}(1-u_1-\ldots-u_k)^{n-r_k-1}du_1du_2\ldots du_k,$$

a result which is independent of $f(x)$. The domain over which this density function is defined is the region for which $u_i \geq 0$ $(i=1,2,..,k)$ and $\sum_1^k u_i \leq 1$.

4.6 Mean Values of Sample Moments when Sample Values are Grouped; Sheppard Corrections

Suppose that x is a continuous random variable having probability element $f(x)dx$, and that O_n is a sample from a population having this distribution. Let the x axis be divided into non-overlapping intervals of equal length δ, suppose I_0 is the interval including the origin, and let h be the x-coordinate of the center of I_0. Denote the intervals by $\ldots, I_{-2}, I_{-1}, I_0, I_1, I_2, \ldots$ where the end points of I_i are $(h+(i-\frac{1}{2})\delta, h+(i+\frac{1}{2})\delta)$, $i = ..,-2,-1,0,1,2,\ldots$ Let

$$(a) \qquad p_i = \int_{h+(i-\frac{1}{2})\delta}^{h+(i+\frac{1}{2})\delta} f(x)dx,$$

the probability associated with I_i. If $f(x)$ is identically zero outside some finite interval there will be only a finite number of non-zero p_i, otherwise there will be a convergent series of p_i. Let n_i be the number of x's in O_n falling into I_i, and let the value of each of these x's be replaced by $h + i\delta$, the midpoint of I_i. Let ${}_\delta M'_r$ be the r-th "grouped" moment of the sample, defined as follows

$$(b) \qquad {}_\delta M'_r = \frac{1}{n}\sum_i n_i (h+i\delta)^r.$$

It will be noted that ${}_\delta M'_r$ is the "grouped" analogue of

$$(c) \qquad M'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r,$$

where $x_1, x_2, \ldots, x_n$ are the values of x in the sample. In fact $M'_r = \operatorname{Lim}_{\delta \to 0} {}_\delta M'_r$. It is easy to verify that $E(M'_r) = \mu'_r$, where

$$(d) \qquad \mu'_r = \int_{-\infty}^{\infty} x^r f(x)\,dx.$$

The problem to be considered here is that of finding $E({}_\delta M'_r)$, where h is a continuous random variable distributed uniformly (i. e. with probability element $\frac{1}{\delta}dh$) on the interval $(-\frac{1}{2}\delta, \frac{1}{2}\delta)$. For a given δ, the random variables involved in the grouping problem are the n_i and h. The conditional probability law of the n_i given h is the multinomial distribution

$$(e) \qquad P = \frac{n!}{\ldots n_{-2}! n_{-1}! n_0! n_1! n_2! \ldots} \ldots p_{-2}^{n_{-2}} p_{-1}^{n_{-1}} p_0^{n_0} p_1^{n_1} p_2^{n_2} \ldots$$

Now we have

$$(f) \qquad E({}_\delta M'_r) = \frac{1}{\delta}\int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \sum {}_\delta M'_r P\,dh,$$

where $\sum$ denotes summation over all positive integral or zero values of the n_i such that $\sum_i n_i = n$. The m. g. f. of ${}_\delta M'_r$ is

$$(g) \qquad \phi(\theta) = E(e^{\theta\, {}_\delta M'_r}) = \frac{1}{\delta}\int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} e^{\theta\, {}_\delta M'_r} P\,dh = \frac{1}{\delta}\int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \Big(\sum_i p_i e^{\theta(h+i\delta)^{r/n}}\Big)^n dh.$$

If the m. g. f. does not exist then the characteristic function (obtained by replacing θ

by $\theta \sqrt{-1}$) will exist since the p_i are positive and will form a convergent series if there is not a finite number of them. We now have

$$(h) \qquad E({}_\delta M_r') = \phi'(0) = \frac{1}{\delta} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \sum_i p_i (h+i\delta)^r dh.$$

Making use of (a) we may write

$$(i) \qquad E({}_\delta M_r') = \frac{1}{\delta} \sum_i \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \int_{h+(i-\frac{1}{2})\delta}^{h+(i+\frac{1}{2})\delta} f(x)(h+i\delta)^r dx\, dh.$$

Setting $h + i\delta = y$, we have

$$(j) \qquad E({}_\delta M_r') = \frac{1}{\delta} \sum_i \int_{\delta(i-\frac{1}{2})}^{\delta(i+\frac{1}{2})} \int_{y-\frac{1}{2}\delta}^{y+\frac{1}{2}\delta} f(x) y^r dx\, dy$$

$$= \frac{1}{\delta} \int_{-\infty}^{\infty} \int_{y-\frac{1}{2}\delta}^{y+\frac{1}{2}\delta} f(x) y^r dx\, dy.$$

Interchanging the order of integration, we obtain

$$(k) \qquad E({}_\delta M_r') = \frac{1}{\delta} \int_{-\infty}^{\infty} \int_{x-\frac{1}{2}\delta}^{x+\frac{1}{2}\delta} f(x) y^r dy\, dx = \frac{1}{\delta(r+1)} \int_{-\infty}^{\infty} [(x+\tfrac{1}{2}\delta)^{r+1} - (x-\tfrac{1}{2}\delta)^{r+1}] f(x) dx.$$

In particular, for r = 1, 2, 3, (k) becomes

$$E({}_\delta M_1') = \int_{-\infty}^{\infty} x f(x) dx = \mu_1',$$

$$E({}_\delta M_2') = \int_{-\infty}^{\infty} (x^2 + \frac{\delta^2}{12}) f(x) dx = \mu_2' + \frac{\delta^2}{12},$$

$$E({}_\delta M_3') = \int_{-\infty}^{\infty} (x^3 + \frac{\delta^2}{4} x) f(x) dx = \mu_3' + \frac{\delta^2}{4} \mu_1'.$$

It will be noted that ${}_\delta M_1'$, $({}_\delta M_2' - \frac{\delta^2}{12})$, and ${}_\delta M_3' - \frac{\delta^2}{4} {}_\delta M_1'$ are <u>unbiased</u> (§6.21) estimates of μ_1', μ_2', μ_3'. The quantities $\frac{\delta^2}{12}$, $-\frac{\delta^2}{4} {}_\delta M_1'$ are called <u>Sheppard corrections</u> of ${}_\delta M_2'$

and ${}_\delta M'_j$. Such corrections can be obtained for higher values of r by further use of (k). Similarly one can determine Sheppard corrections for grouped moments about the sample mean, as defined by

$$ {}_\delta M_r = \frac{1}{n}\sum_i n_i[(h+i\delta) - \frac{1}{n}\sum_j n_j(h+j\delta)]^r. $$

4.7 Appendix on Lagrange's Multipliers

We frequently encounter the problem of finding the extreme (maximum or minimum) value of a function $g(x_1,\ldots,x_n)$ subject to side conditions

(a) $$ \phi_i(x_1,\ldots,x_n) = 0, \qquad i = 1,\ldots,k < n. $$

To insure the independence of the conditions (a) we assume that for some $x_{n_1},\ldots,x_{n_k}$,

$$ \frac{\partial(\phi_1,\ldots,\phi_k)}{\partial(x_{n_1},\ldots,x_{n_k})} \neq 0 $$

at the extremum. To simplify the notation, assume $n_i=i$, $i=1,\ldots,k$. At the extremum, $dg=0$,

(b) $$ \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}\, dx_i = 0, $$

where $dx_1,\ldots,dx_k$ are functions of $dx_{k+1},\ldots,dx_n$, determined by $d\phi_j=0$, i. e.,

(c) $$ \sum_{i=1}^{n} \frac{\partial \phi_j}{\partial x_i}\, dx_i = 0, \qquad j = 1,\ldots,k, $$

and $dx_{k+1},\ldots,dx_n$ are completely arbitrary numbers. In order that (b) be satisfied for all $dx_1,\ldots,dx_n$, which are arbitrary except that they must satisfy (c), a necessary and sufficient condition is that the equation (b) be a linear combination of equations (c), i. e., that for some $\lambda_1,\ldots,\lambda_k$,

(d) $$ \frac{\partial g}{\partial x_i} = -\sum_{j=1}^{k} \lambda_j \frac{\partial \phi_j}{\partial x_i}, \qquad i = 1,\ldots,n. $$

We see that the conditions (d) are obtained if we employ the following rule: To minimize g subject to (a), form the function

$$ G(x_1,\ldots,x_n;\lambda_1,\ldots,\lambda_k) = g + \sum_{j=1}^{k} \lambda_j \phi_j $$

and set

(e) $$ \frac{\partial G}{\partial x_i} = 0, \qquad i = 1,\ldots,n. $$

The equations (a) and (e) constitute a system of n+k equations in n+k unknowns $x_1,\ldots,x_n$; $\lambda_1,\ldots,\lambda_k$. For an extremum it is necessary that $x_1,\ldots,x_n$ satisfy these equations. In most applications in statistics the question of sufficiency can be settled in an obvious way.

CHAPTER V

SAMPLING FROM A NORMAL POPULATION

Since the normal distribution appears in such a wide variety of problems, we shall consider in detail certain sampling problems from such a distribution. Many distributions are important in statistics for the reason that they arise in connection with sampling from a normal universe. In the present chapter, we shall only consider certain sampling problems, deriving certain sampling distribution. The application of these sampling problems to problems of significance tests, statistical estimation, etc., will be made in later chapters.

5.1 Distribution of Sample Mean

An important property of the normal distribution is the so-called reproductive property. We wish to demonstrate that a linear function of normally distributed variates is again normally distributed. Suppose $x_1, x_2, \ldots, x_n$ are distributed independently according to $N(a_1,\sigma_1^2)$, $N(a_2,\sigma_2^2)$, ..., $N(a_n,\sigma_n^2)$, respectively. Let us find the distribution of the linear form $L = l_1x_1 + l_2x_2 + \ldots + l_nx_n$. According to the results of §2.74, the expected value of L is

$$(a)\quad E(l_1x_1+l_2x_2+\ldots+l_nx_n) = l_1E(x_1) + l_2E(x_2) + \ldots + l_nE(x_n) = l_1a_1 + l_2a_2 + \ldots + l_na_n.$$

The joint distribution of the x's is

$$(b)\qquad \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_1\ldots\sigma_n}\, e^{-\frac{1}{2}\left[\frac{(x_1-a_1)^2}{\sigma_1^2}+\ldots+\frac{(x_n-a_n)^2}{\sigma_n^2}\right]}.$$

From this we shall find the moment generating function of the linear form minus its mean, L - E(L),

(c) $$\phi(\theta) = E(e^{\theta[L-E(L)]})$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_1\cdots\sigma_n} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{(x_1-a_1)^2}{\sigma_1^2}+\cdots+\frac{(x_n-a_n)^2}{\sigma_n^2}\right]+\theta[l_1(x_1-a_1)+\cdots+l_n(x_n-a_n)]} dx_1\cdots dx_n$$

$$= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{(x_i-a_i)^2}{\sigma_i^2}+\theta l_i(x_i-a_i)} dx_i$$

$$= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{x_i-a_i+\sigma_i^2\theta l_i}{\sigma_i}\right]^2+\frac{1}{2}\sigma_i^2\theta^2 l_i^2} dx_i$$

$$= e^{\frac{\theta^2}{2}\left(\sum_{i=1}^{n}\sigma_i^2 l_i^2\right)}$$

This is the moment generating function for the probability element

(d) $$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} dy,$$

where

$$\sigma^2 = \sum_{i=1}^{n}\sigma_i^2 l_i^2 .$$

Therefore L is distributed according to $N(\sum_1^n l_i a_i, \sum_1^n l_i^2\sigma_i^2)$. We have the

Theorem (A): If $x_1, x_2, \ldots, x_n$ are independently distributed according to $N(a_1,\sigma_1^2)$, $N(a_2,\sigma_2^2)$, $\ldots, N(a_n\sigma_n^2)$, respectively, then any linear function of the x's $l_1x_1 + l_2x_2 + \cdots + l_nx_n$ is distributed according to

$$N(\sum_1^n l_i a_i, \sum_1^n l_i^2\sigma_i^2).$$

From this result we can easily derive the distribution of the mean of a sample. Consider a sample, O_n, of n observations $x_1, x_2, \ldots, x_n$. The x's are independently distributed each according to $N(a,\sigma^2)$. If we take $l_1 = \frac{1}{n}, \ldots, l_n = \frac{1}{n}$, the linear form L is simply $\bar{x}$, the mean of the sample. Its expected value is

(e) $$\frac{1}{n}a + \frac{1}{n}a + \ldots + \frac{1}{n}a = a;$$

its variance is

(f) $$\frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \ldots + \frac{1}{n^2}\sigma^2 = \frac{1}{n}\sigma^2.$$

Therefore, we have the following corollary to Theorem (A):

Corollary (A_1): If $O_n : x_1, x_2, \ldots, x_n$ is a sample from the normal population $N(a,\sigma^2)$, then the sample mean $\bar{x}$ is distributed according to $N(a,\frac{\sigma^2}{n})$.

5.11 Distribution of Difference between Two Sample Means

Suppose we have two samples, O_n and $O_{n'}$, of n and n' observations drawn from normal populations, $N(a,\sigma^2)$ and $N(a',\sigma'^2)$, respectively. Then the two sample means, $\bar{x}$ and $\bar{x}'$, are distributed according to $N(a, \sigma^2/n)$ and $N(a', \sigma'^2/n')$, respectively. To find the distribution of the difference of the two means, let us consider the linear function $\bar{x} - \bar{x}'$. In this case $l_1 = 1$, $l_2 = -1$; so the expected value of the linear form is

(a) $$a - a',$$

and its variance is

(b) $$\frac{\sigma^2}{n} + \frac{\sigma'^2}{n'}.$$

We therefore have the following corollary to Theorem (A):

Corollary (A_2): If $O_n : x_1, x_2, \ldots, x_n$ and $O_{n'} : x'_1, x'_2, \ldots, x'_{n'}$ are samples from the populations $N(a,\sigma^2)$ and $N(a',\sigma'^2)$, respectively, then $\bar{x} - \bar{x}'$ is distributed according to $N(a-a', \frac{\sigma^2}{n} + \frac{\sigma'^2}{n'})$ where $\bar{x}$ and $\bar{x}'$ are the means of O_n and $O_{n'}$, respectively.

5.12 Joint Distribution of Means in Samples from a Normal Bivariate Distribution

Let us consider a sample $O_n(x_{1\alpha}, x_{2\alpha}, \alpha = 1,2,..,n)$ from the bivariate distribution

(a) $$\frac{\sqrt{A}}{2\pi} e^{-\frac{1}{2}\sum_{i,j=1}^{2} A_{ij}(x_i - a_i)(x_j - a_j)}$$

where $A_{11} = \frac{1}{\sigma_1^2(1-\rho^2)}$, $A_{22} = \frac{1}{\sigma_2^2(1-\rho^2)}$, $A_{12} = \frac{-1}{\sigma_1\sigma_2(1-\rho^2)}$, and $A = |A_{ij}|$.

Let $\bar{x}_i = \frac{1}{n}\sum_{\alpha=1}^{n} x_{i\alpha}$, $i = 1,2$. We wish to determine the joint distribution of $\bar{x}_1$, $\bar{x}_2$. To do this, we determine the m. g. f. of the $(\bar{x}_1 - a_1)$ and $(\bar{x}_2 - a_2)$, i. e.

(b) $$\phi(\theta_1,\theta_2) = E(e^{\sum_1^2 \theta_i(\bar{x}_i-a_i)})$$

$$= (\frac{\sqrt{A}}{2\pi})^n \underbrace{\int_{-\infty}^{\infty}\cdots\int}_{2n\text{-fold}} e^{-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{i,j=1}^{2}A_{ij}(x_{i\alpha}-a_i)(x_{j\alpha}-a_j)+\sum_{\alpha=1}^{n}\sum_{i=1}^{2}\frac{\theta_i}{n}(x_{i\alpha}-a_i)} dx_{11}dx_{21}\cdots dx_{1n}dx_{2n}$$

$$= [\frac{\sqrt{A}}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{i,j=1}^{2}A_{ij}(x_i-a_i)(x_j-a_j)+\sum_{i=1}^{2}\frac{\theta_i}{n}(x_i-a_i)} dx_1dx_2]^n.$$

But, we know from (d) and (e) of §3.22 that if we set $x_i - a_i = y_i$ inside [ ] the resulting expression inside [ ] will be

$$e^{\frac{1}{2n^2}\sum_{i,j=1}^{2}A^{ij}\theta_i\theta_j}.$$

Therefore, the m. g. f. of $(\bar{x}_1 - a_1)$, and $(\bar{x}_2 - a_2)$ is

(c) $$\phi(\theta_1,\theta_2) = e^{\frac{1}{2}\sum_{i,j=1}^{2}(\frac{A^{ij}}{n})\theta_i\theta_j}$$

Since $e^{\frac{1}{2}\sum_{i,j=1}^{2}A^{ij}\theta_i\theta_j}$ is the m. g. f. for (x_1-a_1) and (x_2-a_2) in distribution (b) §3.22, it follows that the distribution of $(\bar{x}_1-a_1)$, $(\bar{x}_2-a_2)$ (having m. g. f. (c)) is

(d) $$\frac{\sqrt{A}n}{2\pi} e^{-\frac{n}{2}\sum_{i,j=1}^{2}A_{ij}(\bar{x}_i-a_i)(\bar{x}_j-a_j)} d\bar{x}_1d\bar{x}_2.$$

We therefore have

Theorem (B): If x_1 and x_2 are distributed jointly according to the normal bivariate law (a) §5.12, then if $\bar{x}_1$ and $\bar{x}_2$ are sample means of the $x_{1\alpha}$ and the $x_{2\alpha}$, respectively, in a sample $O_n(x_{i\alpha}, i=1,2; \alpha=1,2,..,n)$ from such a distribution then $\bar{x}_1$ and $\bar{x}_2$ are also distributed according to a normal bivariate distribution given by (d).

Theorem (B) extends at once to the case of means in a sample from a k-variate normal population with distribution (b) §3.23. The distribution of the means in this case is

(e) $$\frac{n^{k/2}\sqrt{A}}{(2\pi)^{k/2}} e^{-\frac{n}{2}\sum_{i,j=1}^{k}A_{ij}(\bar{x}_i-a_i)(\bar{x}_j-a_j)} d\bar{x}_1d\bar{x}_2\cdots d\bar{x}_k.$$

5.2 The χ^2-distribution

The χ^2-distribution function with m degrees of freedom is defined as

$$f_m(\chi^2)d\chi^2 = \frac{(\frac{\chi^2}{2})^{\frac{m}{2}-1}}{2\Gamma(\frac{m}{2})} e^{-\frac{1}{2}\chi^2} d(\chi^2).$$

This distribution arises very frequently in connection with sampling theory of quadratic forms of normally distributed variables. We shall consider some of the important cases in this chapter and others in Chapters VIII and IX.

The integrals $\int_{\chi_0^2}^{\infty} f_m(\chi^2)d\chi^2$ and $\int_0^{\chi_0^2} f_m(\chi^2)d\chi^2$ are tabulated in many places for various values of m and χ_0^2. When we let $\chi^2/2 = t$, the latter integral is transformed into the Incomplete Gamma Function of which extensive tables have been computed by Karl Pearson.

5.21 Distribution of Sum of Squares of Normally and Independently Distributed Variables

The simplest sample statistic which is distributed according to the χ^2-law is the sum of squares of variates independently distributed according to the same normal law with zero mean. Let us use the method of moment generating functions to find the distribution of $\chi^2 = \sum_1^n x_i^2$ where each x_i (i= 1,2,...,n) is independently distributed according to N(0,1). The joint distribution of the x's is

$$\frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_1^n x_i^2}.$$

Now let us find the moment generating function of χ^2.

(a) $$\phi(\theta) = E(e^{\theta(\sum_1^n x_i^2)}) = \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_1^n x_i^2 + \theta\sum_1^n x_i^2} dx_1 \ldots dx_n$$

$$= \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2\theta)\sum_1^n x_i^2} dx_1 \ldots dx_n$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2\theta)x_i^2} dx_i$$

$$= \left(\frac{1}{\sqrt{1-2\theta}}\right)^n = (1-2\theta)^{-\frac{n}{2}},$$

for $\theta < \frac{1}{2}$.

But this is the moment generating function of the Pearson Type III distribution ((e) of §3.3) when $\beta = \frac{1}{2}$, $\alpha = 0$, and $\nu = \frac{n}{2}$. Therefore by uniqueness Theorem (B), §2.81, we have

Theorem (A): If $O_n : x_1, x_2, \ldots, x_n$ is a sample from $N(0,1)$, the function $\chi^2 = \sum_1^n x_i^2$ is distributed according to the χ^2-law with n degrees of freedom, i. e.

$$\text{(b)} \qquad \frac{1}{2} \frac{(\chi^2)^{\frac{n}{2}-1}}{\Gamma(\frac{n}{2})} e^{-\frac{1}{2}\chi^2} d(\chi^2).$$

From this result it follows that, if $x_1, x_2, \ldots, x_n$ are distributed independently according to $N(a,\sigma^2)$, then $\chi^2 = \sum_1^n (x_i-a)^2/\sigma^2$ is distributed according to $f_n(\chi^2)d(\chi^2)$.

We can readily determine the moments of the χ^2 distribution from its moment generating function. We expand $\phi(\theta)$ in a power series

$$\text{(c)} \qquad \phi(\theta) = 1 + \frac{n}{2}.2\theta + \frac{\frac{n}{2}(\frac{n}{2}+1)}{2!}(2\theta)^2 + \ldots + \frac{\frac{n}{2}(\frac{n}{2}+1)\ldots(\frac{n}{2}+h-1)}{h!}(2\theta)^h + \ldots$$

Then we find the moments about zero

$$\text{(d)} \qquad E[(\chi^2)^h] = \frac{\partial^h \phi}{\partial \theta^h}\bigg|_{\theta=0} = 2^h \cdot \frac{n}{2}(\frac{n}{2}+1)\ldots(\frac{n}{2}+h-1).$$

The mean is n and the variance is

$$\sigma^2 = n(n+2) - n^2 = 2n.$$

5.22 Distribution of the Exponent in a Multivariate Normal Distribution

Now let us consider a normal multivariate distribution of k variates with zero means

$$\text{(a)} \qquad \frac{\sqrt{|A|}}{(2\pi)^{k/2}} e^{-\frac{1}{2}\sum_{1,j=1}^{k} A_{ij}x_i x_j},$$

and let us find the distribution of the quadratic form, $\sum_{1,j=1}^{k} A_{ij}x_i x_j$. To do this we find the moment generating function of the quadratic form,

$$\text{(b)} \quad \phi(\theta) = E(e^{\theta\sum_{1}^{k} A_{ij}x_i x_j}) = \frac{\sqrt{|A|}}{(2\pi)^{k/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{1}^{k} A_{ij}x_i x_j + \theta\sum_{1}^{k} A_{ij}x_i x_j} dx_1 \ldots dx_k$$

$$= \frac{\sqrt{|A|}}{(2\pi)^{k/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{1}^{k} (1-2\theta)A_{ij}x_i x_j} dx_1 \ldots dx_k .$$

It follows from §3.23 that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{1}^{k} B_{ij}x_i x_j} dx_1 \ldots dx_k = \frac{(2\pi)^{k/2}}{\sqrt{|B|}};$$

the above integration yields

$$\phi(\theta) = \frac{\sqrt{|A|}}{(2\pi)^{k/2}} \cdot \frac{(2\pi)^{k/2}}{\sqrt{|(1-2\theta)A_{ij}|}} = \frac{\sqrt{|A|}}{\sqrt{(1-2\theta)^k}\sqrt{|A|}}.$$

That is,

$$\text{(c)} \qquad \phi(\theta) = (1-2\theta)^{-\frac{k}{2}} \qquad (\theta < \tfrac{1}{2}),$$

which, as will be seen from (e) in §3.3, is the m. g. f. for a χ^2 distribution with k degrees of freedom.

We therefore have

Theorem (A): If $x_1, x_2, .., x_k$ are distributed according to the normal multivariate law (a), then $\sum_{i,j=1}^{k} A_{ij}x_i x_j = \chi^2$ say, is distributed according to $f_k(\chi^2)$.

More generally, the quadratic form $\sum_{i,j,=1}^{k} A_{ij}(x_i - a_i)(x_j - a_j)$ from the distribution

$$\frac{\sqrt{|A|}}{(2\pi)^{k/2}} e^{-\frac{1}{2}\sum_{1}^{k} A_{ij}(x_i - a_i)(x_j - a_j)},$$

has the χ^2 distribution with k degrees of freedom.

5.23 Reproductive Property of χ^2-Distribution

In the same way that the normal distribution possesses the reproductive property so also does the χ^2-distribution. Suppose we have $\chi_1^2, \chi_2^2, \ldots, \chi_k^2$ distributed according to $f_{m_1}(\chi_1^2), f_{m_2}(\chi_2^2), \ldots, f_{m_k}(\chi_k^2)$, respectively. From the joint distribution of these variates, let us find the moment generating function of the sum $\sum_{i=1}^{k} \chi_i^2$, assuming independence,

$$\phi(\theta) = E(e^{\theta \sum_1^k \chi_i^2}) = \int_0^\infty \cdots \int_0^\infty e^{\theta \sum_1^k \chi_i^2} f_{m_1}(\chi_1^2) \cdots f_{m_k}(\chi_k^2)\, d\chi_1^2 \cdots d\chi_k^2$$

$$= \prod_{i=1}^{k} \int_0^\infty e^{\theta \chi_i^2} f_{m_i}(\chi_i^2) d\chi_i^2$$

$$= \prod_{i=1}^{k} [(1-2\theta)^{-\frac{m_i}{2}}]$$

$$= (1-2\theta)^{-\frac{m}{2}},$$

where $m = \sum_{i=1}^{k} m_i$. $\phi(\theta)$ is the m. g. f. for a χ^2-distribution with m degrees of freedom. Therefore, we have the following

Theorem (A): If $\chi_1^2, \chi_2^2, \ldots, \chi_k^2$ are independently distributed according to χ^2-laws with $m_1, m_2, \ldots, m_k$ degrees of freedom respectively, then $\sum_1^k \chi_i^2$ is distributed according to a χ^2-law with $\sum_1^k m_i$ degrees of freedom.

5.24 Cochran's Theorem

Cochran's theorem states certain conditions under which a set of quadratic forms are independently distributed according to χ^2-laws if the variables of the quadratic forms are independently distributed, each according to N(0,1). To prove this theorem, we need several algebraic theorems which will be stated as lemmas.

Lemma 1: If q is a quadratic form, $\sum_{\alpha,\beta=1}^{n} a_{\alpha\beta} x_\alpha x_\beta$, of order n and rank r, there exists a linear transformation $z_\alpha = \sum_{\beta=1}^{n} b_{\alpha\beta} x_\beta (\alpha = 1, 2, \ldots, r)$ such that $\sum_{\alpha,\beta=1}^{n} a_{\alpha\beta} x_\alpha x_\beta = \sum_{\alpha=1}^{r} c_\alpha z_\alpha^2$, where the c_α are +1 or -1.

In §3.23 we exhibited a linear transformation that would do this for a positive definite quadratic form. The reader may extend that demonstration to prove Lemma 1.*

*A proof of Lemma 1 is given in M. Bôcher, Introduction to Higher Algebra, Macmillan, New York, 1907.

Lemma 2: If $\sum_{\alpha,\beta=1}^{n} A_{\alpha\beta}x_\alpha x_\beta$ is transformed into $\sum_{\alpha,\beta=1}^{n} a'_{\alpha\beta}z_\alpha z_\beta$ by a linear transformation, $z_\alpha = \sum_{\beta=1}^{n} b_{\alpha\beta}x_\beta (\alpha=1,2,\ldots,n)$ then

$$|a_{\alpha\beta}| = |a'_{\alpha\beta}| \cdot |b_{\alpha\beta}|^2.$$

This lemma can be readily verified from the fact that $a_{\alpha\beta} = \sum_{\gamma,\delta=1}^{n} a'_{\gamma\delta}b_{\gamma\alpha}b_{\delta\beta}$ and by using the rule for multiplying determinants.

Lemma 3: Suppose we have k quadratic forms, $q_1, q_2,\ldots,q_k$, in $x_1, x_2,\ldots,x_n$ of ranks $n_1, n_2,\ldots,n_k$, respectively, and suppose $\sum_{1=1}^{k} q_1 = \sum_{\alpha=1}^{n} x_\alpha^2$. Then a necessary and sufficient condition that there exist a non-singular linear transformation $z_\alpha = \sum_{\beta=1}^{n} c_{\alpha\beta}x_\beta$ $(\alpha=1,2,\ldots,\sum_{1}^{k} n_1)$ such that

$$q_1 = z_1^2 + \ldots + z_{n_1}^2,$$

$$- - - - - - - - - - -$$

$$q_k = z_{n_1+\ldots+n_{k-1}+1}^2 + \ldots + z_{n_1+\ldots+n_k}^2$$

is that $n = n_1 + n_2 + \ldots + n_k$.

Proof: The necessity condition is obvious since $\sum_{1}^{k} n_1$ must be equal to n in order for the transformation to be non-singular.

Now consider the sufficiency condition. We assume $n = n_1 + n_2 + \ldots + n_k$. By Lemma 1 there is a linear transformation $y_\alpha^{(1)} = \sum_{\beta=1}^{n} b_{\alpha\beta}^{(1)}x_\beta$ such that $q_1 = \sum_{\alpha=1}^{n_1} c_\alpha (y_\alpha^{(1)})^2$, where $c_\alpha = +1$ or -1. In the same way we know there exist transformations

$$y_\alpha^{(2)} = \sum_{\beta=1}^{n} b_{\alpha\beta}^{(2)}x_\beta, \ \ldots, \ y_\alpha^{(k)} = \sum_{\beta=1}^{n} b_{\alpha\beta}^{(k)}x_\beta,$$

such that

$$q_2 = \sum_{\alpha=n_1+1}^{n_1+n_2} c_\alpha (y_\alpha^{(2)})^2, \ \ldots, \ q_k = \sum_{\alpha=n_1+\ldots+n_{k-1}+1}^{n} c_\alpha (y_\alpha^{(k)})^2.$$

In other words we have n_1 linear forms $y_\alpha^{(1)} = \sum_{\beta=1}^{n} b_{\alpha\beta}^{(1)}x_\beta (\alpha=1,\ldots,n_1)$ such that $q_1 = \sum_{\alpha=1}^{n_1} c_\alpha (\sum_{\beta=1}^{n} b_{\alpha\beta}^{(1)}x_\beta)^2 = \sum_{\alpha=1}^{n_1} c_\alpha (y_\alpha^{(1)})^2$; we have n_2 linear forms such that

$$q_2 = \sum_{\alpha=n_1+1}^{n_1+n_2} c_\alpha (\sum_{\beta} b_{\alpha\beta}^{(2)}x_\beta)^2 = \sum_{\alpha=n_1+1}^{n_1+n_2} c_\alpha (y_\alpha^{(2)})^2, \text{ etc.}$$

Let us denote $y_\alpha^{(1)}$ by z_α for $\alpha = 1,2,\ldots,n_1$, $y_\alpha^{(2)}$ by z_α for $\alpha = n_1+1, \ldots, n_1 + n_2$, etc. Let us denote $b_{\alpha\beta}^{(1)}$ by $c_{\alpha\beta}$ for $\alpha = 1,\ldots,n_1$ $(\beta=1,\ldots,n)$, $b_{\alpha\beta}^{(2)}$ by $c_{\alpha\beta}$ for $\alpha = n_1+1, \ldots,$

n_1+n_2 ($\beta=1,\ldots,n$), etc. Combining all of the linear transformations, we may write

$$z_\alpha = \sum_{\beta=1}^{n} c_{\alpha\beta}x_\beta \qquad (\alpha = 1,2,\ldots,n).$$

Then $q_1 = \sum_{\alpha=1}^{n_1} c_\alpha z_\alpha^2$, $q_2 = \sum_{\alpha=n_1+1}^{n_1+n_2} c_\alpha z_\alpha^2$, etc., and $\sum_{\alpha=1}^{n} x_\alpha^2 = \sum_{i=1}^{k} q_i = \sum_{\alpha=1}^{n} c_\alpha z_\alpha^2 = \sum_{\alpha=1}^{n} c_\alpha (\sum_{\beta=1}^{n} c_{\alpha\beta}x_\beta)^2$.

By Lemma 2, $|\delta_{\alpha\beta}| = |c_\alpha \delta_{\alpha\beta}| \cdot |c_{\alpha\beta}|^2$ (where $\delta_{\alpha\beta}$ is 1 if $\alpha = \beta$ and is 0 if $\alpha \neq \beta$). This reduces to

$$1 = (\prod_{\alpha=1}^{n} c_\alpha).|c_{\alpha\beta}|^2.$$

Since the $c_\alpha = \pm 1$, this equation is

$$1 = \pm 1 \cdot |c_{\alpha\beta}|^2,$$

and because the $c_{\alpha\beta}$ are real $|c_{\alpha\beta}| = \pm 1$.

This fact tells us that the n linear forms are independent and constitute a non-singular linear transformation. From the identity

$$\sum x_\alpha^2 = \sum c_\alpha z_\alpha^2$$

we deduce that $\sum c_\alpha z_\alpha^2$ is positive definite since $\sum x_\alpha^2$ is positive definite. Hence, each $c_\alpha = +1$. This proves the sufficiency of the condition $n = n_1 + n_2 + \ldots + n_k$. It is interesting to observe that $|c_{\alpha\beta}| = +1$ and that $\sum_{\gamma=1}^{n} c_{\alpha\gamma}c_{\beta\gamma} = \delta_{\alpha\beta}$, that is, the transformation is orthogonal.

Cochran's theorem follows readily from this algebraic theorem.

Theorem (A) (Cochran's Theorem): If $x_\alpha(\alpha=1,2,\ldots,n)$ are independently distributed according to $N(0,1)$ and if $\sum_{\alpha=1}^{n} x_\alpha^2 = \sum_{i=1}^{k} q_i$ where q_i is a quadratic form of rank n_i, a necessary and sufficient condition that the q_i be distributed according to $f_{n_i}(\chi^2)$ is that $\sum_{i=1}^{k} n_i = n$.

Proof: Assume $n_i = n$, and find the m. g. f. of the q_i. We have

$$\phi = E(e^{\sum_1^k \theta_i q_i}) = \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_1^n x_\alpha^2 + \sum_1^k \theta_i q_i} \prod_1^n dx_\alpha.$$

Now transform the x's to z's by Lemma 3, noting that the Jacobian is unity.

$$\phi = \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_1^n z_\alpha^2 + \theta_1(z_1^2+\cdots+z_{n_1}^2)+\cdots+\theta_k(z_{n_1}^2+\cdots+n_{k-1}+1^2+\cdots+z_n^2)} \prod_1^n dz_\alpha$$

$$= \prod_{i=1}^{k} (1-2\theta_i)^{-\frac{n_i}{2}},$$

which is the m. g. f. of k independent χ^2 distributions with $n_1, n_2, \ldots, n_k$ degrees of freedom, thus establishing the sufficiency condition.

The converse assumes that

$$\prod_{i=1}^{k} (1-2\theta_i)^{-\frac{n_i}{2}} = \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_1^n x_\alpha^2 + \sum_1^k \theta_i q_i} \prod_1^n dx_\alpha.$$

Since $\sum_{i=1}^{k} q_i = \sum_{\alpha=1}^{n} x_\alpha^2$, the right hand side of the equation becomes the m. g. f. of $\sum_{\alpha=1}^{n} x_\alpha^2$ (which has a χ^2 distribution) when $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$. So the equation becomes

$$\prod_{i=1}^{k} (1-2\theta)^{-\frac{n_i}{2}} = (1-2\theta)^{-\frac{n}{2}},$$

that is,

$$(1-2\theta)^{-\sum_1^k \frac{n_i}{2}} = (1-2\theta)^{-\frac{n}{2}}.$$

Hence, $\sum_1^k n_i = n$, and the theorem is proved.

5.25 Independence of Mean and Sum of Squared Deviations from Mean in Samples From a Normal Population

As an application of Cochran's Theorem, we shall show that the sample mean and sum of squares of deviations about the mean in a sample from a normal population are independent and have χ^2-distributions. Consider a sample $O_n: x_1, x_2, \ldots, x_n$ drawn from a normal population $N(0,1)$. Then

(a) $$\sum_{\alpha=1}^{n} x_\alpha^2 = \sum_{\alpha=1}^{n} (x_\alpha - \bar{x})^2 + n\bar{x}^2.$$

Let

$$q_1 = \sum_{\alpha=1}^{n} (x_\alpha - \bar{x})^2 = \sum_{\alpha=1}^{n} (x_\alpha - \frac{1}{n}\sum_1^n x_\beta)(x_\alpha - \frac{1}{n}\sum_1^n x_\gamma) = \sum_{\alpha,\beta=1}^{n} (\delta_{\alpha\beta} - \frac{1}{n}) x_\alpha x_\beta,$$

and

$$q_2 = n\bar{x}^2 = n(\tfrac{1}{n}\textstyle\sum x_\alpha)(\tfrac{1}{n}\textstyle\sum x_\beta) = \displaystyle\sum_{\alpha,\beta=1}^{n} \tfrac{1}{n} x_\alpha x_\beta .$$

q_2 is of rank 1, for in the matrix

$$\left\| \begin{matrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \cdots & \cdots & \cdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{matrix} \right\|$$

any minor of order two,

$$\begin{vmatrix} \frac{1}{n} & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} \end{vmatrix},$$

is zero, but each element is different from zero. The determinant of the matrix of q_1 is

$$D = \begin{vmatrix} 1-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} \cdots & -\frac{1}{n} \\ \cdot & \cdot & \cdot \cdots \cdot & \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} \cdots & 1-\frac{1}{n} \end{vmatrix} .$$

Subtracting the first row from each of the others, we get

$$D = \begin{vmatrix} 1-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} \cdots & -\frac{1}{n} \\ -1 & 1 & 0 \cdots & 0 \\ -1 & 0 & 1 \cdots & 0 \\ \cdot & \cdot & \cdot \cdots & \cdot \\ -1 & 0 & 0 \cdots & 1 \end{vmatrix} .$$

Next we add each column to the first and find

$$D = \begin{vmatrix} 1-\frac{n}{n} & -\frac{1}{n} & -\frac{1}{n} \cdots & -\frac{1}{n} \\ 0 & 1 & 0 \cdots & 0 \\ 0 & 0 & 1 \cdots & 0 \\ \cdot & \cdot & \cdot \cdots & \cdot \\ 0 & 0 & 0 \cdots & 1 \end{vmatrix} = 0,$$

for all the elements of the first column are zero. If we use this method of evaluation on any principal minor of order $n - 1$, we get

$$M = \begin{vmatrix} 1 - \frac{n-1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{vmatrix} = \frac{1}{n} \neq 0.$$

Hence the rank of q_1 is $n - 1$. Using Cochran's Theorem we conclude that $\sum(x_\alpha - \bar{x})^2$ and $n\bar{x}^2$ are independently distributed according to $f_{n-1}(\chi^2)$ and $f_1(\chi^2)$, respectively.

If x is distributed according to $N(a,\sigma^2)$ then $\frac{x-a}{\sigma}$ is distributed according to $N(0,1)$. Hence, we have proved the following corollary to Cochran's Theorem:

If $O_n: x_1, \ldots, x_n$ is a sample from $N(a,\sigma^2)$, then $\sum_1^n \frac{(x_\alpha - \bar{x})^2}{\sigma^2}$ and $n\frac{(\bar{x}-a)^2}{\sigma^2}$ are independently distributed according to $f_{n-1}(\chi^2)$ and $f_1(\chi^2)$. It also follows that $s^2 = \sum_1^n \frac{(x_\alpha - \bar{x})^2}{n-1}$ and $\bar{x}$ are independently distributed.

It should be pointed out that one could establish the fact that $\sum_1^n \frac{(x_\alpha - \bar{x})^2}{\sigma^2}$ and $\frac{(\bar{x}-a)}{\sigma}\sqrt{n}$ for a sample from $N(a,\sigma^2)$ are independently distributed according to $f_{n-1}(\chi^2)$ and $N(0,1)$, respectively, by verifying that the m. g. f.

$$\phi(\theta_1,\theta_2) = E\left(e^{\theta_1 \frac{\sum_1^n (x_\alpha - \bar{x})^2}{\sigma^2} + \frac{\theta_2(\bar{x}-a)\sqrt{n}}{\sigma}}\right) = (1-2\theta_1)^{-\frac{n-1}{2}} \cdot e^{\frac{1}{2}\theta_2^2}.$$

5.3 The "Student" t-Distribution

Next we shall derive the distribution of the ratio of two independent variates, one normally distributed and the other distributed according to the χ^2-law. Let ξ be a variate distributed according to $N(0,1)$ and let χ^2 be distributed according to $f_m(\chi^2)$. If these are independently distributed, the joint probability element is

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} \frac{(\frac{\chi^2}{2})^{\frac{m}{2}-1}}{2\Gamma(\frac{m}{2})} e^{-\frac{1}{2}\chi^2} d\xi\, d(\chi^2).$$

Let us change variables to

$$t = \frac{\xi}{\sqrt{\frac{\chi^2}{m}}}, \quad u = \chi^2, \qquad -\infty < t < \infty, \quad 0 < u < \infty.$$

Then $\xi = t\sqrt{\frac{u}{m}}$, $\chi^2 = u$.

The Jacobian of this transformation is

$$J = \begin{vmatrix} \sqrt{\frac{u}{m}} & \frac{t}{2\sqrt{um}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{u}{m}}.$$

Hence the joint distribution of t and u is

$$\frac{1}{2\sqrt{2\pi}\,\Gamma(\frac{m}{2})} (\tfrac{u}{2})^{\frac{m}{2}-1} e^{-\frac{1}{2}(t^2\frac{u}{m}+u)} \sqrt{\tfrac{u}{m}}\, du\, dt.$$

To find the marginal distribution of t, we integrate out u,

$$g_m(t) = \frac{1}{2\sqrt{m}\sqrt{2\pi}\,\Gamma(\frac{m}{2})} \int_0^\infty \sqrt{2}\, (\tfrac{u}{2})^{\frac{m}{2}-\frac{1}{2}} e^{-(\frac{t^2}{m}+1)\frac{u}{2}}\, du$$

$$= \frac{(1+\frac{t^2}{m})^{-\frac{m+1}{2}}}{2\sqrt{m}\sqrt{\pi}\,\Gamma(\frac{m}{2})} \int_0^\infty [(\tfrac{t^2}{m}+1)\tfrac{u}{2}]^{\frac{m+1}{2}-1} e^{-(\frac{t^2}{m}+1)\frac{u}{2}} (\tfrac{t^2}{m}+1)\, du,$$

(a) $$g_m(t) = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})\sqrt{m}\sqrt{\pi}} (1+\tfrac{t^2}{m})^{-\frac{m+1}{2}}.$$

$g_m(t)$ is called the "Student" t-distribution with m degrees of freedom. Values of t_ϵ have been tabulated such that

$$\int_{-t_\epsilon}^{t_\epsilon} g_m(t)dt = \epsilon,$$

for $\epsilon = .1, .2, .3, .4, .5, .6, .7, .8, .9, .95, .98, .99$ and $m = 1, 2, 3, \ldots, 30$, in R. A. Fisher's Statistical Methods for Research Workers.

The application of this distribution to sampling theory is immediate. As an important application consider a sample O_n from $N(a,\sigma^2)$. Then

$$\xi = \frac{(\bar{x}-a)\sqrt{n}}{\sigma}$$

is distributed according to $N(0,1)$ and

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{\sigma^2},$$

is independently distributed according to $f_{n-1}(\chi^2)$. The ratio

$$t = \frac{\xi}{\sqrt{\frac{\chi^2}{n-1}}} = \frac{\frac{(\bar{x}-a)\sqrt{n}}{\sigma}}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{\sigma^2(n-1)}}} = \frac{(\bar{x}-a)\sqrt{n(n-1)}}{\sqrt{\sum(x_i-\bar{x})^2}} = \frac{(\bar{x}-a)\sqrt{n}}{s}$$

is, therefore, distributed according to $g_{n-1}(t)$.

The quantity t and its sampling theory which marked a new step in statistical inference were first investigated by Gossett who without rigorously proving his result suggested the above distribution of t in a paper published in 1908 under the name of "Student". A rigorous proof was supplied by R. A. Fisher in 1926. The essential feature of t is that both it and its distribution are functionally independent of σ.

The "Student" distribution may also be used in connection with two samples. Let $O_{n_1}(x_{1\alpha}, \alpha = 1,2,\ldots,n_1)$ and $O_{n_2}(x_{2\alpha}, \alpha = 1,2,\ldots,n_2)$ be samples from $N(a_1,\sigma^2)$ and $N(a_2,\sigma^2)$, respectively. Let $\bar{x}_1$ and $\bar{x}_2$ be sample means and $s_1^2 = \sum_{\alpha=1}^{n_1}(x_{1\alpha}-\bar{x}_1)^2/(n_1-1)$ and $s_2^2 = \sum_{\alpha=1}^{n_2}(x_{2\alpha}-\bar{x}_2)^2/(n_2-1)$. Then

$$\xi = \frac{(\bar{x}_1-\bar{x}_2)-(a_1-a_2)}{\sqrt{\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})}}$$

is distributed according to $N(0,1)$ and

$$\chi^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{\sigma^2}$$

is distributed independently according to $f_{n_1+n_2-2}(\chi^2)$. Hence, the ratio

$$t = \frac{\xi}{\sqrt{\frac{\chi^2}{n_1+n_2-2}}} = \frac{(\bar{x}_1-\bar{x}_2)-(a_1-a_2)}{\sqrt{(n_1-1)s_1^2+(n_2-1)s_2^2}} \frac{\sqrt{n_1+n_2-2}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

is distributed according to $g_{n_1+n_2-2}(t)$.

It can be verified by the reader that

$$\lim_{m \to \infty} g_m(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

5.4 Snedecor's F-Distribution

Now let us consider the distribution of the ratio of two quantities independently distributed according to χ^2-distributions. Let χ_1^2 and χ_2^2 be independently distributed according to $f_{m_1}(\chi_1^2)$ and $f_{m_2}(\chi_2^2)$, respectively. The joint distribution is

$$\frac{1}{2\Gamma(\frac{m_1}{2})} (\frac{\chi_1^2}{2})^{\frac{m_1}{2}-1} e^{-\frac{\chi_1^2}{2}} \cdot \frac{1}{2\Gamma(\frac{m_2}{2})} (\frac{\chi_2^2}{2})^{\frac{m_1}{2}-1} e^{-\frac{\chi_2^2}{2}} .$$

Let us make the change of variables

$$F = \frac{\chi_1^2}{m_1} \Big/ \frac{\chi_2^2}{m_2}, \qquad v = \chi_2^2, \qquad \begin{matrix} 0 < F < \infty, \\ 0 < v < \infty. \end{matrix}$$

Then

$$\chi_1^2 = v \frac{m_1}{m_2} F \qquad \chi_2^2 = v,$$

and the Jacobian of the transformation is

$$J = \begin{vmatrix} v\frac{m_1}{m_2} & \frac{m_1}{m_2}F \\ 0 & 1 \end{vmatrix} = \frac{m_1}{m_2} v .$$

The distribution of the transformed variates is

$$\frac{(\frac{m_1}{m_2})^{\frac{m_1}{2}}}{2\Gamma(\frac{m_1}{2})\Gamma(\frac{m_2}{2})} F^{\frac{m_1}{2}-1} (\frac{v}{2})^{\frac{m_1}{2}+\frac{m_2}{2}-1} e^{-\frac{1}{2}(\frac{m_1}{m_2}Fv+v)} .$$

Integrating out the extraneous variable v, we get the distribution of F,

$$\frac{(\frac{m_1}{m_2})^{\frac{m_1}{2}} F^{\frac{m_1}{2}-1}}{\Gamma(\frac{m_1}{2})\Gamma(\frac{m_2}{2})} (1+\frac{m_1}{m_2}F)^{-\frac{m_1+m_2}{2}} \int_0^\infty [(1+\frac{m_1}{m_2}F)\frac{v}{2}]^{\frac{m_1+m_2}{2}-1} e^{-(1+\frac{m_1}{m_2}F)\frac{v}{2}} dv$$

(a) $$= \frac{\Gamma(\frac{m_1+m_2}{2})}{\Gamma(\frac{m_1}{2})\Gamma(\frac{m_2}{2})} (\frac{m_1}{m_2})^{\frac{m_1}{2}} F^{\frac{m_1}{2}-1} (1 + \frac{m_1}{m_2} F)^{-\frac{m_1+m_2}{2}} .$$

This distribution, known as Snedecor's F-distribution with m_1 and m_2 degrees of freedom, will be denoted by $h_{m_1,m_2}(F)$.

Values of F_ϵ have been tabulated such that

$$\int_0^{F_\epsilon} h_{m_1,m_2}(F)dF = \epsilon,$$

for $\epsilon = .99, .95$ and all combinations of (m_1,m_2) from (1,1) to (12,30) and for certain combinations from (14,32) to (500,1000), in Snedecor's Statistical Methods.

The moments about zero are easily obtained. Since the above is a distribution function, the integral over the entire range of F is unity, and, hence,

$$\int_0^\infty F^{\frac{r_1}{2}-1} (1 + \frac{m_1}{m_2} F)^{-\frac{r_1+r_2}{2}} dF = \frac{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})}{\Gamma(\frac{r_1+r_2}{2})} (\frac{m_2}{m_1})^{\frac{r_1}{2}} .$$

Using this fact, we get μ'_r by integration.

(b) $$E(F^r) = \frac{\Gamma(\frac{m_1+m_2}{2})}{\Gamma(\frac{m_1}{2})\Gamma(\frac{m_2}{2})} (\frac{m_1}{m_2})^{\frac{m_1}{2}} \int_0^\infty F^{\frac{m_1}{2}-1+r} (1 + \frac{m_1}{m_2} F)^{-\frac{m_1+m_2}{2}} dF .$$

$$= \frac{\Gamma(\frac{m_1+m_2}{2})}{\Gamma(\frac{m_1}{2})\Gamma(\frac{m_2}{2})} (\frac{m_1}{m_2})^{\frac{m_1}{2}} \frac{\Gamma(\frac{m_1}{2}+r)\Gamma(\frac{m_2}{2}-r)}{\Gamma(\frac{m_1+m_2}{2})} (\frac{m_2}{m_1})^{\frac{m_1}{2}+r}$$

$$= \frac{\Gamma(\frac{m_1}{2}+r)\Gamma(\frac{m_2}{2}-r)}{\Gamma(\frac{m_1}{2})\Gamma(\frac{m_2}{2})} (\frac{m_2}{m_1})^r ,$$

for $r < \frac{m_2}{2}$.

By a simple change of variable the F-distribution may be changed into a Type I distribution, (the integrand of the Beta function times a constant). Let

$$x = \frac{\frac{m_1}{m_2} F}{1 + \frac{m_1}{m_2} F} .$$

Then

$$F = \frac{m_2}{m_1} \frac{x}{1 - x} ,$$

and

$$dF = \frac{m_2}{m_1} \frac{dx}{(1 - x)^2} .$$

So $h_{m_1,m_2}(F)dF$ transforms into

$$(c) \quad \frac{\Gamma(\frac{m_1+m_2}{2})}{\Gamma(\frac{m_1}{2})\Gamma(\frac{m_2}{2})} (\frac{m_1}{m_2})^{\frac{m_1}{2}} x^{\frac{m_1+m_2}{2}} (\frac{m_2}{m_1} \cdot \frac{x}{1-x})^{-\frac{m_1}{2}-1} \frac{m_2}{m_1} \frac{dx}{(1-x)^2} = \frac{1}{B(\frac{m_1}{2},\frac{m_2}{2})} x^{\frac{m_2}{2}-1} (1-x)^{\frac{m_1}{2}-1} dx .$$

It should be pointed out that the square of Student's t is simply distributed as $h_{1,m}(t^2)d(t^2)$.

If we make the change of variable

$$(d) \quad z = \frac{1}{2} \log_e F ,$$

we obtain R. A. Fisher's z-distribution.

Example 1: As an example of the applications of the F-distribution, consider two samples $O_{n_1}:(x_{1\alpha}, \alpha = 1,2,\ldots,n_1)$ and $O_{n_2}:(x_{2\alpha} \alpha = 1,2,\ldots,n_2)$ from populations $N(a_1,\sigma_1^2)$ and $N(a_2,\sigma_2^2)$, respectively. Let

$$s_1^2 = \sum_{\alpha=1}^{n_1} \frac{(x_{1\alpha}-\bar{x}_1)^2}{n_1 - 1} \text{ and } s_2^2 = \sum_{\alpha=1}^{n_2} \frac{(x_{2\alpha}-\bar{x}_2)^2}{n_2 - 1} .$$

Then

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = (\frac{s_1^2}{s_2^2})(\frac{\sigma_2^2}{\sigma_1^2})$$

is distributed according to $h_{n_1-1,n_2-1}(F)$.

Example 2: Suppose O_{n_1}, O_{n_2},...,O_{n_k} are k samples from $N(a_1,\sigma^2)$, $N(a_2,\sigma^2)$,..., $N(a_k,\sigma^2)$, respectively. Then

$$\sum_1^k \frac{(n_1-1)s_1^2}{\sigma^2}$$

has a χ^2-distribution with $n - k$ ($\sum n_i = n$) degrees of freedom. Since $\frac{n_1(\bar{x}_1-a_1)^2}{\sigma^2}$ is distributed as $f_1(\chi^2)$, the sum

$$\sum_{i=1}^{k} \frac{n_i(\bar{x}_i-a_i)^2}{\sigma^2},$$

is distributed according to $f_k(\chi^2)$ and independently of the former sum. From these facts it follows that the ratio

$$F = \frac{\sum_1^k n_i(\bar{x}_i-a_i)^2}{\sum_1^k (n_i-1)s_i^2} \cdot \frac{n-k}{k}$$

is distributed according to $h_{k,n-k}(F)$.

If the k samples come from a normal populations with the same mean as well as the same variance,

$$\frac{\sqrt{n_i}(\bar{x}_i-a)}{\sigma} \qquad (i = 1,2,\ldots,k)$$

is distributed according to $N(0,1)$ and, therefore,

$$\sum_{i=1}^{k} \frac{n_i(\bar{x}_i-\bar{x})^2}{\sigma^2} \qquad \left(\bar{x} = \frac{\sum_1^k n_i\bar{x}_i}{\sum_1^k n_i}\right)$$

has the χ^2-distribution with $k - 1$ degrees of freedom. From this fact, and since the $\bar{x}_i$ are distributed independently of the s_i^2, it follows that

$$F = \frac{\sum_1^k n_i(\bar{x}_i-\bar{x})^2}{\sum_1^k (n_i-1)s_i^2} \cdot \frac{n-k}{k-1}$$

has the distribution $h_{k-1,n-k}(F)$.

5.5 Distribution of Second Order Sample Moments in Samples from a Bivariate Normal Distribution

Let us consider a sample 0_n ($x_{1\alpha}$, $x_{2\alpha}$, $\alpha = 1,2,\ldots,n$) from the distribution (a) in §5.12. The probability density function for the sample is

$$\frac{|A|^{\frac{n}{2}}}{(2\pi)^n} e^{-\frac{1}{2}\sum_{i,j=1}^{2}\sum_{\alpha=1}^{n}(x_{i\alpha}-a_i)(x_{j\alpha}-a_j)}$$

We shall find the m. g. f. of the sample variance and twice the covariance

$$\phi(\theta_{11},\theta_{12},\theta_{22}) = E(e^{\sum\theta_{ij}a_{ij}}),$$

where $a_{ij} = \sum_{\alpha=1}^{n}(x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j)$ and $\bar{x}_i = \frac{1}{n}\sum_{\alpha=1}^{n} x_{i\alpha}$, $\theta_{ij} = \theta_{ji}$. We have

$$\phi = \frac{|A|^{\frac{n}{2}}}{(2\pi)^n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{i,j,\alpha} A_{ij}(x_{i\alpha}-a_i)(x_{j\alpha}-a_j) + \sum_{i,j,\alpha} \theta_{ij}(x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j)} \prod_{\alpha} dx_{1\alpha} dx_{2\alpha}$$

$$= \frac{|A|^{\frac{n}{2}}}{(2\pi)^n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}Q} \prod_{\alpha} dx_{1\alpha} dx_{2\alpha},$$

where

$$Q = \sum_{i,j,\alpha} A_{ij}(x_{i\alpha}-a_i)(x_{j\alpha}-a_j) - 2\sum_{i,j,\alpha} \theta_{ij}(x_{i\alpha}-a_i)(x_{j\alpha}-a_j) + 2\left(n\sum_{i,j} \theta_{ij}(\bar{x}_i-a_i)(\bar{x}_j-a_j)\right).$$

$$\phi = \frac{|A|^{\frac{n}{2}}}{(2\pi)^n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{i,j,\alpha,\beta} B_{ij,\alpha\beta}(x_{i\alpha}-a_i)(x_{j\beta}-a_j)} \prod_{\alpha} dx_{1\alpha} dx_{2\alpha}$$

$$= \frac{|A|^{\frac{n}{2}}}{|B|^{\frac{1}{2}}}.$$

The determinant of the matrix of the quadratic form $\sum B_{ij,\alpha\beta}(x_{i\alpha}-a_i)(x_{j\beta}-a_j)$ is of order 2n and is

$$|B_{ij,\alpha\beta}| = \begin{vmatrix} C & D & D & \cdots & D \\ D & C & D & \cdots & D \\ D & D & C & \cdots & D \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ D & D & D & \cdots & C \end{vmatrix}$$

where C is a 2 x 2 block of elements as follows:

$$\begin{matrix} A_{11} - 2\theta_{11}(1 - \frac{1}{n}) & A_{12} - 2\theta_{12}(1 - \frac{1}{n}) \\ A_{21} - 2\theta_{21}(1 - \frac{1}{n}) & A_{22} - 2\theta_{22}(1 - \frac{1}{n}) \end{matrix}$$

and D is a 2 x 2 block of elements as follows:

$$\begin{matrix} \frac{2}{n}\theta_{11} & \frac{2}{n}\theta_{12} \\ \frac{2}{n}\theta_{21} & \frac{2}{n}\theta_{22} \end{matrix}.$$

If in $|B_{ij,\alpha\beta}|$ the first row of elements is subtracted from the third, fifth, etc., and

the second row is subtracted from the fourth, sixth, etc., and if in the resulting determinant to the first column is added the third, fifth, etc., and to the second column is added the fourth, sixth, etc., we find that

$$|B_{ij,\alpha\beta}| = |A_{ij}| \cdot |A_{ij} - 2\theta_{ij}|^{n-1},$$

which exists if the θ_{ij} are sufficiently small. Hence the m. g. f. of the a_{11}, a_{22} and $2a_{12}$ is

$$\text{(a)} \qquad \phi(\theta_{11},\theta_{12},\theta_{22}) = |A_{ij}|^{\frac{n-1}{2}} |A_{ij}-2\theta_{ij}|^{-\frac{n-1}{2}}.$$

Now if we can find a function $f(a_{11},a_{12},a_{22})$ such that

$$\text{(b)} \qquad \phi(\theta_{11},\theta_{12},\theta_{22}) = \iiint_R e^{\sum \theta_{ij}a_{ij}} f(a_{11},a_{12},a_{22})da_{11}da_{12}da_{22},$$

where R is the region in the space of the a_{ij} for which $a_{11} > 0$, $a_{22} > 0$, $-1 < \frac{a_{12}}{\sqrt{a_{11}a_{22}}} < 1$, then $f(a_{11},a_{12},a_{22})$ will be the p. d. f. of the a_{ij}. The uniqueness of the solution can be argued from the multivariate analogue of Theorem (B) of §2.81.

Denoting $|A_{ij}|^{(n-1)/2}$ by $A^{(n-1)/2}$ and $A_{ij} - 2\theta_{ij}$ by $\bar{A}_{ij}$, and choosing values of the θ_{ij} small enough for $||\bar{A}_{ij}||$ to be positive definite, we can write

$$\text{(c)} \qquad |\bar{A}_{ij}|^{-\frac{n-1}{2}} = \bar{A}_{11}^{-(\frac{n-1}{2})} \bar{A}_{22}^{-(\frac{n-1}{2})} (1-k^2)^{-\frac{n-1}{2}}, \qquad k = \bar{A}_{12} \Big/ \sqrt{\bar{A}_{11}\bar{A}_{22}},$$

and we can expand $(1-k^2)^{-(n-1)/2}$ into the infinite series $[1/\Gamma((n-1)/2)]\sum_{i=0}^{\infty}[\Gamma((n-1)/2 + i)/i!]k^{2i}$. Hence we may write

$$\text{(d)} \qquad \phi(\theta_{11},\theta_{12},\theta_{22}) = \frac{A^{\frac{n-1}{2}}}{\Gamma(\frac{n-1}{2})} \sum_{i=0}^{\infty} \frac{\Gamma(\frac{n-1}{2} + i)\, \bar{A}_{12}^{2i}}{i!\, \bar{A}_{11}^{\frac{n-1}{2}+i} \bar{A}_{22}^{\frac{n-1}{2}+i}}.$$

But

$$\bar{A}_{11}^{-(\frac{n-1}{2}+1)} = \int_0^{\infty} \frac{(\frac{a_{11}}{2})^{\frac{n-1}{2}+1-1}}{\Gamma(\frac{n-1}{2}+1)} e^{-\frac{1}{2}\bar{A}_{11}a_{11}} d(\tfrac{a_{11}}{2}),$$

and a similar expression holds for $\bar{A}_{22}^{-((n-1)/2+1)}$. Therefore, we may write

$$(e)\quad \phi(a_{11},a_{12},a_{22}) = \int_0^\infty \int_0^\infty \frac{A^{\frac{n-1}{2}}}{\Gamma(\frac{n-1}{2})}(\frac{a_{11}a_{22}}{4})^{\frac{n-3}{2}} e^{-\frac{1}{2}A_{11}a_{11}-\frac{1}{2}A_{22}a_{22}}\left[\sum_{i=0}^{\infty}\frac{(\frac{a_{11}a_{22}}{4})^i A_{12}^{2i}}{i!\,(\frac{n-1}{2}+i)}\right]d(\frac{a_{11}}{2})d(\frac{a_{22}}{2}).$$

If in [ ] we make use of the formula $i! = (2i)!\sqrt{\pi}/2^{2i}\Gamma(i+\frac{1}{2})$ we may write [ ] as

$$(f)\qquad \sum_{i=0}^{\infty}\frac{(a_{11}a_{22})^i A_{12}^{2i}\,\Gamma(i+\frac{1}{2})}{\sqrt{\pi}(2i)!\,\Gamma(\frac{n-1}{2}+i)}.$$

But from the definition of the Beta function, §3.3,

$$\frac{\Gamma(i+\frac{1}{2})}{\Gamma(\frac{n-1}{2}+i)} = \frac{1}{\Gamma(\frac{n-2}{2})}\int_0^1 t^{i-\frac{1}{2}}(1-t)^{\frac{n-4}{2}}dt$$

$$= \frac{1}{\Gamma(\frac{n-2}{2})}\int_{-1}^{1} r^{2i}(1-r^2)^{\frac{n-4}{2}}dr.$$

Therefore

$$[\ ] = \frac{1}{\sqrt{\pi}\Gamma(\frac{n-2}{2})}\int_{-1}^{1}\sum_{i=0}^{\infty}\frac{(\sqrt{a_{11}a_{22}})^{2i}A_{12}^{2i}}{(2i)!}r^{2i}(1-r^2)^{\frac{n-4}{2}}dr$$

$$= \frac{1}{\sqrt{\pi}\Gamma(\frac{n-2}{2})}\int_{-1}^{1}\sum_{j=0}^{\infty}\frac{(\sqrt{a_{11}a_{22}})^{j}A_{12}^{j}}{j!}r^{j}(1-r^2)^{\frac{n-4}{2}}dr = \frac{1}{\sqrt{\pi}\Gamma(\frac{n-2}{2})}\int_{-1}^{1} e^{r\sqrt{a_{11}a_{22}}A_{12}}(1-r^2)^{\frac{n-4}{2}}dr,$$

since terms for odd values of j vanish upon integration. Making use of this value of [ ] in (e) we have

$$(g)\quad \phi(a_{11},a_{12},a_{22}) = \int_0^\infty\int_0^\infty\int_{-1}^{1}\frac{A^{\frac{n-1}{2}}(\frac{1}{2})^{n-1}}{\sqrt{\pi}\Gamma(\frac{n-1}{2})\Gamma(\frac{n-2}{2})}(a_{11}a_{22})^{\frac{n-3}{2}}(1-r^2)^{\frac{n-4}{2}}$$

$$e^{-\frac{1}{2}A_{11}a_{11}-\frac{1}{2}A_{22}a_{22}+r\sqrt{a_{11}a_{22}}A_{12}}da_{11}da_{22}dr.$$

Setting $r\sqrt{a_{11}a_{22}} = a_{12}$, (g) can be expressed as (b) where

$$(h)\qquad f(a_{11},a_{12},a_{22}) = \frac{A^{\frac{n-1}{2}}\,|a_{ij}|^{\frac{n-4}{2}}}{2^{n-1}\sqrt{\pi}\Gamma(\frac{n-1}{2})\Gamma(\frac{n-2}{2})}e^{-\frac{1}{2}\sum A_{ij}a_{ij}}.$$

As we mentioned earlier, the uniqueness of this p. d. f. may be argued from the multivariate analogue of Theorem (B) §2.81.

The sampling distribution of the correlation coefficient r may be found by setting $a_{12} = r\sqrt{a_{11}a_{22}}$ in (h), expanding $e^{r\sqrt{a_{11}a_{12}}A_{12}}$ into an infinite series, and integrating with respect to a_{11} and a_{22}; we obtain as the probability element of r

$$(1) \qquad f(r)dr = \frac{(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}}}{\sqrt{\pi}\,\Gamma(\frac{n-1}{2})\Gamma(\frac{n-2}{2})}\sum_{i=0}^{\infty}\frac{(2\rho r)^i}{i!}\Gamma^2(\frac{n-1+i}{2})\,dr,$$

where $\rho = -(A_{12}/\sqrt{A_{11}A_{22}})$, the correlation coefficient of the population.

If $\rho = 0$, the distribution of r is simply

$$(j) \qquad f(r)dr = \frac{\Gamma(\frac{n-1}{2})}{\sqrt{\pi}\,\Gamma(\frac{n-2}{2})}(1-r^2)^{\frac{n-4}{2}}\,dr.$$

The distribution (h) may be generalized to the case of a sample from a k-variate normal distribution given by (b) in §3.23. The distribution for the k-variate case, which will be derived in Chapter XI, is

$$(k) \qquad f(a_{ij}) = \frac{A^{\frac{n-1}{2}}|a_{ij}|^{\frac{n-k-2}{2}}e^{-\frac{1}{2}\sum_{i,j=1}^{k}A_{ij}a_{ij}}}{2^{\frac{(n-1)k}{2}}\pi^{\frac{k(k-1)}{4}}\Gamma(\frac{n-1}{2})\Gamma(\frac{n-2}{2})\cdots\Gamma(\frac{n-k}{2})},$$

where $a_{ij} = \sum_{\alpha=1}^{n}(x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j)$, $x_{i\alpha}(i = 1,2,\ldots,k;\ \alpha = 1,2,\ldots,n)$ being the sample. Clearly, $n > k$ for this distribution to exist.

This is a very important distribution function and is fundamental in the theory of normal multivariate statistical analysis. It is known as the Wishart distribution.

5.6 Independence of Second Order Moments and Means in Samples from a Normal Multivariate Distribution

In §5.25 it was shown that in samples of size n from a normal distribution $N(a,\sigma^2)$, the quantities $(1/\sigma^2)\sum_{\alpha=1}^{n}(x_\alpha-\bar{x})^2$ and $\frac{n(\bar{x}-a)}{\sigma}$ were independently distributed according to $f_{n-1}(\chi^2)$ and $N(0,1)$, respectively.

In the case of samples of size n from the k-variate normal distribution (b),

§3.23, the two sets of quantities $a_{ij} = \sum_{\alpha=1}^{n}(x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j)(i,j = 1,2,\ldots,k)$ and $\bar{x}_i (i = 1,2)$ are independently distributed according to (k), §5.5, and (e), §5.12, respectively. A straight forward method of establishing the independence of the two systems is by evaluating the characteristic function of a_{ii} and $2a_{ij}$ $(i\neq j)$, and $(\bar{x}_i-a_i)\sqrt{n}$:

$$\phi(\theta_{ij},\theta_i) = E(e^{\sum_1^k \theta_{ij}a_{ij} + \sum_1^k \theta_i(\bar{x}_i-a_i)\sqrt{n}}),$$

where $\theta_{ij} = \theta_{ji}$, which turns out to be a product of the form $\phi_1(\theta_{ij}) \cdot \phi_2(\theta_i)$, i. e.

$$A^{\frac{n-1}{2}}|A_{ij}-2\theta_{ij}|^{-\frac{n-1}{2}} \cdot e^{\frac{1}{2}\sum_1^k A^{ij}\theta_i\theta_j}.$$

CHAPTER VI

ON THE THEORY OF STATISTICAL ESTIMATION

Let O_n be a sample from a population whose c.d.f. depends on h parameters θ_1, $\theta_2,\ldots,\theta_h$. Suppose the functional form of the c.d.f. is known, but the true values of the parameters are unknown. A fundamental problem in the theory of statistical estimation is the following: On the basis of the evidence of O_n, can we assign an interval for one of the parameters, say θ_1, and then state with a given amount of confidence (the meaning of this phrase will have to be defined) that the true value of θ_1 lies in this interval? More generally, can we make similar statements regarding a subset of the parameters, say $\theta_1,\theta_2,\ldots,\theta_m$ $m \leq h$, and a region in the parameter space These problems are discussed in §6.1. If instead of assigning on the basis of O_n an interval of values in which we estimate the true parameter value to be contained, we wish to assign a single value, the problem is more difficult: We can hardly hope that our "point estimate" will coincide exactly with the true value: In what sense can such an estimate be said to be "good"? How can "good" estimates be found? These questions are considered in §6.2. Closely related to the problem of point estimation of one or more parameters are questions of curve fitting; these are taken up in §6.4.

The problems described above may be called parametric problems in statistical estimation. There are also non-parametric cases of statistical estimation. One of these is the problem of tolerance limits, which may be formulated as follows: Suppose a sample O_n is from a population in which the random variable x is continuous. Can we determine functions L_1 and L_2 of the x's in the sample such that we can state with a given probability that $100\beta\%$ of the x's in the population will be included in the interval (L_1,L_2), no matter what the population distribution is? or no matter what the values of the parameters are if the functional form of the distribution is known? This problem is discussed in §6.3. Some of the underlying sampling theory is discussed in §4.55.

6.1 Confidence Intervals and Confidence Regions

In this section we consider the estimation of one or more parameters by means of statements that the parameter lies, or the parameters lie, in a certain region of the parameter space. The discussion of the example of §6.11 should be carefully studied: while this will not be repeated elsewhere, the analogous considerations pertain in every case taken up in §§6.11-6.13.

6.11 Case in which the Distribution Depends on Only One Parameter

It will be clearest if we begin by means of an example (range of a rectangular distribution): Let R be the range of a sample O_n from a population with the p. d. f.

$$f(x;\theta) = 1/\theta, \text{ when } 0 \leq x \leq \theta, \text{ and } 0, \text{ otherwise.}$$

It has been shown in §4.54 that the p. d. f. of R is

$$f_n(R;\theta) = n(n-1)\theta^{-n}R^{n-2}(\theta-R), \qquad 0 \leq R \leq \theta.$$

If we introduce the function

$$\psi = R/\theta,$$

we find that the distribution of this function of sample and parameter is independent of the true value of the parameter, its p. d. f. is

$$g(\psi) = n(n-1)\psi^{n-2}(1-\psi), \qquad 0 \leq \psi \leq 1.$$

We pick a positive number $\epsilon < 1$ (it is customary to take $\epsilon = .95$ or $.99$) and define ψ_ϵ from

$$\int_{\psi_\epsilon}^{1} g(\psi)d\psi = \epsilon.$$

Then regardless of the true value of θ,

$$\epsilon = \Pr(\psi_\epsilon \leq \psi \leq 1) = \Pr(\psi_\epsilon \leq R/\theta \leq 1),$$

which is equivalent to the statement

(a) $$\Pr(R \leq \theta \leq R/\psi_\epsilon) = \epsilon.$$

It should be noted that R is the random variable in this statement and not θ. The interval $\delta:(R,R/\psi_\epsilon)$ is called a confidence interval for θ, and ϵ is called the confidence coefficient. Let us examine the significance of the probability statement (a):

First of all, (a) does not mean that if we take the value of R from a specific sample, say $R = R_1$, that the probability that

(b) $$R_1 \leq \theta \leq R_1/\psi_\epsilon$$

is ϵ: For, θ is not a random variable, it is a constant, even if unknown, and hence the statement (b) is true or false; if (b) is true the probability is unity, and if false,

zero, -- in no case is it ϵ. The situation is analogous to the random drawing (with replacement) of balls from the classical urn, in which the proportion of white balls is ϵ, of black balls, $1 - \epsilon$. After we have drawn a ball the randomness of the process is over, the particular ball drawn is either black or white, and probability statements, aside from the trivial one that $p = 0$ or 1, are no longer possible. However, if we draw a large number of balls we may expect that the percentage of white balls drawn will closely approximate $100\,\epsilon$. More precisely: The law of large numbers (§3.11) tells us that the proportion of white balls drawn converges stochastically to ϵ as the number of drawings is increased.

We now see the practical significance of the probability statement (a): If we always use confidence coefficient ϵ and always assert that the true value of the parameter θ (it need not always be the same parameter) lies in the interval obtained by putting the sample values into the confidence interval, then in the long run (i. e. in repeated sampling) the percentage of correct statements can be expected to be very close to $100\,\epsilon$. Again more precisely, we should say that the probability that the proportion of correct statements departs from ϵ by more than a fixed amount $h > 0$, approaches zero as the number of statements (i.e. number of samples) is increased, no matter how small h.

In general, if a distribution depends on one parameter θ, and if we have two functions $\underline{\theta}(O_n)$ and $\overline{\theta}(O_n)$ which depend on the sample O_n but not on θ, so that the interval

$$\delta(O_n) : \underline{\theta}(O_n) \leqq \theta \leqq \overline{\theta}(O_n)$$

is a random interval, then if the probability that the random interval δ cover the true value of the parameter is ϵ

$$\Pr\{\theta \in \delta(O_n)\} = \epsilon,$$

whatever be the true value θ, we call $\delta(O_n)$ a <u>confidence interval</u> for θ, and ϵ the <u>confidence coefficient</u>. We shall sometimes refer to the pair $\underline{\theta}$, $\overline{\theta}$ of random variables as <u>confidence limits</u>. This terminology is due to Neyman.

The method of finding confidence intervals that was employed in the example is worth noting: It depends on finding a function ψ of O_n and θ whose distribution is independent of θ. If the function ψ is monotone and continuous in θ, then the relation

$$\Pr(\psi_{1\epsilon} \leqq \psi \leqq \psi_{2\epsilon}) = \epsilon \tag{c}$$

can be inverted to read

$$\Pr(\theta \in \delta(O_n)) = \epsilon,$$

where $\delta(O_n)$ is the confidence interval. Another perhaps more direct method of determining confidence intervals is as follows:

Suppose $T(x_1,\ldots,x_n)$ is a function of a sample $O_n : (x_1,x_2,\ldots,x_n)$ from a population with distribution element $f(x;\theta)dx$, such that the probability element of T is $g(T;\theta)dT$. Suppose the range* of values of T having non-zero probability density is (a,b), and suppose the range of possible values of θ is (α,β). Suppose two continuous monotone increasing functions $T_\epsilon(\theta)$ and $T'_\epsilon(\theta)$ exist such that

$$\text{(d)} \qquad \begin{aligned} &\int_a^{T_\epsilon} g(T;\theta)dT = p(1-\epsilon), \\ &\int_{T'_\epsilon}^{b} g(T;\theta)dT = q(1-\epsilon), \end{aligned}$$

where p and q are positive such that $p + q = 1$. Assume that $g(T;\theta)$ is such that T_ϵ and T'_ϵ each ranges from a to b as θ ranges from α to β. Then for a given value of T, let $\underline{\theta}$ and $\overline{\theta}$ be the values of θ for which $T_\epsilon(\theta) = T$, $T'_\epsilon(\theta) = T$, respectively. Then $(\underline{\theta},\overline{\theta})$ is a confidence interval for θ with confidence coefficient ϵ. That $(\underline{\theta},\overline{\theta})$ is a confidence interval follows from the relation

$$\Pr(T_\epsilon(\theta) \leqq T \leqq T'_\epsilon(\theta)) = \epsilon,$$

which, because of the continuous monotonic character of $T_\epsilon(\theta)$ and $T'_\epsilon(\theta)$, may be inverted and written as $\Pr(\underline{\theta} \leqq \theta \leqq \overline{\theta}) = \epsilon$. It should be noted that we may obtain confidence limits for each value of p if functions $T_\epsilon(\theta)$ and $T'_\epsilon(\theta)$ of the required kind exist for each p. The question arises as to which value of p is "best". This would depend, of course, on what definition of "best" we choose. In those cases where the mean value of the length of the confidence interval is a function which factors in the form $h_1(p)h_2(\theta)$, common sense suggests that we should choose p so that the mean length is a minimum. In the case of large samples, the definition of "best" confidence intervals is fairly direct (see §6.12).

We may represent confidence intervals obtained by this process, graphically as follows:

---

*We permit a or α to be $-\infty$, b or β to be $+\infty$.

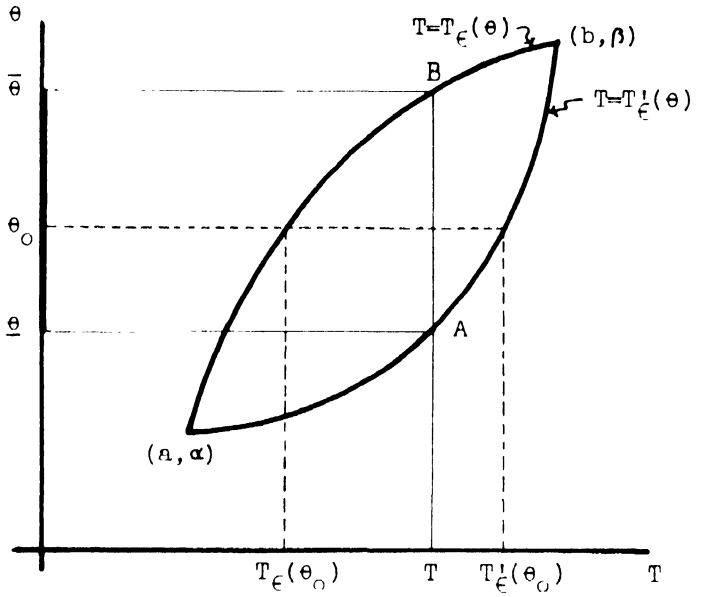

Figure 6

Suppose the true value of θ is θ_0. For any sample value of T, the corresponding confidence interval is formed as follows: Draw a line parallel to the θ-axis, defined by T = sample value. Let A, B be the points of intersection of this line with the two curves as indicated in Fig. 6 . The confidence interval is the projection of the segment AB on the θ-axis. The confidence interval will cover the true value of θ_0 if and only if the segment AB crosses the line $\theta = \theta_0$, that is if and only if T falls in the range $T_\epsilon(\theta_0)$, $T'_\epsilon(\theta_0)$. But the probability of T falling in this interval is precisely ε. We thus have $Pr(\underline{\theta} \leqq \theta_0 \leqq \bar{\theta}) = \epsilon$. The discussion and conclusion hold for any θ_0 in the range (α,β).

This method, for example, has been applied by R. A. Fisher to the problem of determining the confidence limits for ρ from the distribution (1) §5.5 of the correlation coefficient r. Fisher uses the term fiducial limits instead of confidence limits.☆

The idea involved in this method has also been applied to cases where T is a discrete random variable to obtain approximate confidence limits for the parameter involved. In this case the = signs in the analogue of (d) for the discrete case are replaced placed by ≥ signs, and the largest value of $T_\epsilon(\theta)$ and smallest value of $T'_\epsilon(\theta)$ are obtained satisfying the inequalities. $T_\epsilon(\theta)$ and $T'_\epsilon(\theta)$ will be step-functions and the approximate confidence limits are obtained by drawing a smooth curve through the graphs of the step-functions. For example, Clopper and Pearson (Biometrika, Vol. 26 (1934), pp. 404-413) have applied the method to the problem of determining approximate confidence limits for the binomial probability parameter p from the statistic $\frac{x}{n}$ in the binomial distribution

${}_nC_x p^x q^{n-x}$ (x=0,1,2,...,n), and Ricker (Journal of the American Statistical Association, Vol. 32 (1937), pp. 349-356) has applied the method to the Poisson distribution $\frac{m^x}{x!} e^{-m}$, where m is the parameter and x the statistic. A method of determining confidence limits for θ from large samples based on the likelihood function is given in §6.12.

6.12 Confidence Limits from Large Samples

Suppose x has c. d. f. $F(x,\theta)$, where θ is a parameter. Let O_n be a sample of size n from a population having this c. d. f. Let $P(O_n,\theta)$ be the likelihood function, i. e.

$$P(O_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta), \tag{a}$$

where $f(x,\theta)$ is the p. d. f. if x is a continuous variable, and is simply probability* if x is a discrete variable.

We recall the first method of obtaining confidence intervals given in §6.11, which depends on finding a function ψ of O_n and θ whose distribution is independent of θ. That a function of the desired type for large samples may be obtained from the likelihood function $P(O_n,\theta)$ may be concluded by use of the central limit Theorem (C) of §4.21. The central limit theorem applies to a sum (the average), so we replace the product in (a) by a sum by taking logarithms:

$$\log P(O_n,\theta) = \sum_{i=1}^{n} y_i, \tag{b}$$

where $y_i = \log f(x_i,\theta)$ may be regarded as a random variable for any fixed θ. To apply the central limit theorem we need E(y) and σ_y^2, where $y = \log f(x,\theta)$. Now

$$E(y) = \int_{-\infty}^{+\infty} \log f(x,\theta) d_x F(x,\theta), \tag{c}$$

where $d_x F(x,\theta) = f(x,\theta)dx$ in the continuous case, and the integral (c) becomes a sum in th discrete case. The calculation (c) does not give a simple result, but it is clear that if we employed $z = \partial y/\partial\theta$, then

$$E(z) = \int_{-\infty}^{+\infty} \frac{\partial}{\partial\theta} \log f(x,\theta) d_x F(x,\theta),$$

and in the continuous case this becomes

$$E(z) = \int_{-\infty}^{+\infty} \frac{\partial}{\partial\theta} f(x,\theta)dx.$$

*If x is discrete, $f(x,\theta) = F(x,\theta) - F(x-0,\theta)$

If the order of integration and differentiation may be interchanged,

(d) $$E(z) = 0.$$

Let us now assume (d) to be true in any case and furthermore that $A^2 = E(z^2)$ is finite. Differentiating (b) we get

$$\sum_{i=1}^{n} z_i = \frac{\partial}{\partial \theta} \log P(O_n,\theta),$$

and hence

$$\bar{z} = \frac{1}{n} \frac{\partial}{\partial \theta} \log P(O_n,\theta).$$

Applying the central limit theorem to $\bar{z}$ we have that

$$\frac{1}{\sqrt{n}A} \frac{\partial}{\partial \theta} \log P(O_n,\theta)$$

is asymptotically distributed according to N(0,1). We summarize in

Theorem(A): If

$$E\{\frac{\partial}{\partial \theta} \log f(x,\theta)\} = 0, \text{ and } A^2 = E\{[\frac{\partial}{\partial \theta} \log f(x,\theta)]^2\}$$

is finite, then

$$\frac{1}{\sqrt{n}A} \frac{\partial}{\partial \theta} \log P(O_n,\theta)$$

is asymptotically distributed according to N(0,1).

Hence we have approximately, for large n,

(e) $$Pr(-d_\epsilon \leq \frac{1}{\sqrt{n}A} \frac{\partial \log P}{\partial \theta} \leq d_\epsilon) = \epsilon,$$

where d_ϵ is chosen so that

$$\frac{1}{\sqrt{2\pi}} \int_{-d_\epsilon}^{d_\epsilon} e^{-\frac{1}{2}y^2} dy = \epsilon.$$

Now if $\frac{1}{\sqrt{n}A} \frac{\partial \log P}{\partial \theta}$ is monotone in θ, we may invert in (e) and write the result

(f) $$\Pr(\underline{\theta} \leqq \theta \leqq \overline{\theta}) = \epsilon.$$

The asymptotic confidence intervals (f) furnished by $\frac{1}{\sqrt{nA}} \frac{\partial \log P}{\partial \theta}$ are optimum in the following sense: The mean value of $\{\frac{\partial}{\partial \theta} (\frac{1}{\sqrt{nA}} \frac{\partial \log P}{\partial \theta})\}^2$ is greater than that of any other function $G(O_n)$ of the sample which has $N(0,1)$ as its limiting distribution*. This maximum property of the mean squared rate of change with respect to θ implies shortest average confidence intervals in a certain sense, since confidence intervals are obtained by taking the inverse of $\frac{1}{\sqrt{nA}} \frac{\partial \log P}{\partial \theta}$ with respect to θ.

Example: Suppose samples of size n are drawn from a population having the binomial distribution

$$f(x,p) = dF(x,p) = p^x(1-p)^{1-x}, \qquad x = 0,1.$$

In a sample of size n

$$P(O_n,p) = p^{\sum_1^n x_i} (1-p)^{n-\sum_1^n x_i}$$

$$= p^{n_1}(1-p)^{n-n_1},$$

where $n_1 = \sum_1^n x_i$. We verify that $E(\partial \log f/\partial p) = 0$ and calculate

$$A^2 = E\{\frac{\partial \log p^x(1-p)^{1-x}}{\partial p})^2\} = \frac{1}{p(1-p)},$$

and

$$\frac{1}{\sqrt{nA}} \frac{\partial \log P}{\partial p} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} (\frac{n_1}{p} - \frac{n-n_1}{1-p})$$

$$= \frac{n_1 - np}{\sqrt{n}\sqrt{p(1-p)}} = \frac{(\frac{n_1}{n} - p)\sqrt{n}}{\sqrt{p(1-p)}}.$$

Therefore, to find approximate confidence limits with confidence coefficient ϵ, we invert the expression

$$\Pr(-d_\epsilon \leqq \frac{(\frac{n_1}{n} - p)\sqrt{n}}{\sqrt{p(1-p)}} \leqq d_\epsilon) = \epsilon,$$

---

*For proof for the case where $G(O_n)$ is of the form $\sum_1^n h(x_i,\theta)$ where $G(O_n)$ is asymptotically distributed according to $N(0,1)$, see S. S. Wilks, Annals of Math. Stat., Vol. 9 (1938), pp. 166-175, and for more general results, see A. Wald, Annals of Math. Stat., Vol. 13, (1942), pp. 127-137.

obtaining

$$\Pr(\underline{p} \leqq p \leqq \overline{p}) = \epsilon,$$

where $\underline{p}$ and $\overline{p}$ are given as the roots of the quadratic $(\frac{n_1}{n} - p)^2 n = d_\epsilon^2(p - p^2)$.

6.13 Confidence Intervals in the Case where the Distribution Depends on Several Parameters

Suppose that the c. d. f. of the population depends on parameters $\theta_1, \theta_2, \ldots, \theta_h$, and we wish to estimate θ_1. If there exist functions $\underline{\theta}_1(O_n)$, $\overline{\theta}_1(O_n)$ of the sample, such that the probability that the random interval

$$\delta(O_n): \quad \underline{\theta}_1(O_n) \leqq \theta \leqq \overline{\theta}_1(O_n),$$

cover the true value of θ_1 does not depend on the true values of $\theta_1, \theta_2, \ldots, \theta_h$,

$$\Pr\{\theta_1 \in \delta(O_n)\} = \epsilon, \text{ independent of } \theta_1, \theta_2, \ldots, \theta_h,$$

then we say that $\delta(O_n)$ is a confidence interval for θ_1 with confidence coefficient ϵ. (The parameters $\theta_2, \theta_3, \ldots, \theta_h$ are sometimes called nuisance parameters.)

Example 1 (Mean of a normal population): If O_n is a sample from a population with distribution $N(a,\sigma^2)$, then in the notation of §5.3,

$$t = \sqrt{n}(\bar{x}-a)/s$$

has the t-distribution $g_{n-1}(t)$ with n-1 degrees of freedom. Define t_ϵ from

$$\int_{-t_\epsilon}^{t_\epsilon} g_{n-1}(t)dt = \epsilon.$$

Then

$$\epsilon = \Pr(-t_\epsilon \leqq t \leqq t_\epsilon) = \Pr(\bar{x}-t_\epsilon s/\sqrt{n} \leqq a \leqq \bar{x}+t_\epsilon s/\sqrt{n}),$$

whatever be the true values of a and σ^2. Hence $(\bar{x}-t_\epsilon s/\sqrt{n}, \bar{x}+t_\epsilon s/\sqrt{n})$ is a confidence interval for a with confidence coefficient ϵ.

Example 2 (Difference of means of two normal populations known to have the same variance): Let $O_{n_i}:(x_{i1}, x_{i2}, \ldots, x_{in_i})$ be a sample of size n_i from $N(a_i,\sigma^2)$, $i=1,2$. Let

$$s_i = \sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2, \qquad \bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i,$$

$$d = \bar{x}_1-\bar{x}_2, \qquad a = a_1-a_2.$$

Then by §5.25 S_1/σ^2 has the χ^2-distribution with n_1-1 degrees of freedom, hence* (§5.23) S/σ^2, where $S = S_1 + S_2$, has the χ^2-distribution with $n_1 + n_2 - 2$ degrees of freedom. Furthermore, $y = (d-a)/[\sigma^2(n_1^{-1}+n_2^{-1})]^{1/2}$ has the distribution $N(0,1)$, and since y and S/σ^2 are statistically independent (§5.25), it follows from §5.3 that $\sigma y/[S/(n_1+n_2-2)]^{1/2}$ has the t-distribution with $n_1 + n_2 - 2$ degrees of freedom, $g_{n_1+n_2-2}(t)$. Defining t_ϵ from

$$\int_{-t_\epsilon}^{t_\epsilon} g_{n_1+n_2-2}(t)dt = \epsilon,$$

we find by the method of Example 1 that a confidence interval for $\underline{a}$ is $(d-t_\epsilon s', d+t_\epsilon s')$, where

$$s' = \left[\frac{(n_1+n_2)S}{(n_1+n_2-2)n_1n_2}\right]^{\frac{1}{2}}.$$

Example 3 (Variance of a normal distribution): Let O_n be a sample from $N(a,\sigma^2)$. Let

$$S = \sum_{i=1}^{n}(x_i-\bar{x})^2, \qquad \bar{x} = \sum_{i=1}^{n}x_i/n.$$

Then (§5.25) S/σ^2 has the χ^2-distribution $f_{n-1}(\chi^2)$ with n-1 degrees of freedom. Let $\chi^2_{\epsilon 1}$, $\chi^2_{\epsilon 2}$ be any two points on the range $(0,\infty)$ such that

$$\int_{\chi^2_{\epsilon 1}}^{\chi^2_{\epsilon 2}} f_{n-1}(\chi^2)d\chi^2 = \epsilon.$$

We find that $(S/\chi^2_{\epsilon 2}, S/\chi^2_{\epsilon 1})$ is a confidence interval for σ^2 with confidence coefficient ϵ.

Example 4 (Ratio of the variances of two normal distributions): Let O_{n_i}: $(x_{i1},x_{i2},\ldots,x_{in_i})$ be a sample of size n_i from $N(a_i,\sigma_i^2)$, $i=1,2$. Let

$$s_i^2 = \sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2/(n_i-1), \qquad \bar{x}_i = \sum_{j=1}^{n_i}x_{ij}/n_i,$$

$$T = s_1^2/s_2^2, \qquad \theta = \sigma_1^2/\sigma_2^2.$$

Since $(n_i-1)s_i^2$ for $i=1,2$, are independently distributed according to χ^2-distributions with n_i-1 degrees of freedom respectively, it follows from §5.4 that T/θ has the F-distribution $h_{n_1-1,n_2-1}(F)$ with n_1-1 and n_2-1 degrees of freedom. Pick a pair of limits $F_{1\epsilon}$, $F_{2\epsilon}$ so that

*since S_1 and S_2 are statistically independent

$$\int_{F_{1\epsilon}}^{F_{2\epsilon}} h_{n_1-1,n_2-1}(F)dF = \epsilon.$$

Then a confidence interval for θ is $(T/F_{2\epsilon}, T/F_{1\epsilon})$.

6.14 Confidence Regions

We suppose that the population distribution depends on parameters $\theta_1,\theta_2,\ldots,\theta_h$. We denote the parameter point $(\theta_1,\theta_2,\ldots,\theta_h)$ by Θ, and the entire h-dimensional space of admissible parameter values by Ω. If $\delta(O_n)$ is a random region in Ω which depends on the sample O_n, but not on the unknown parameter point Θ, and if the probability that the random region $\delta(O_n)$ cover the true parameter point Θ is independent of Θ,

$$\Pr\{\Theta\epsilon\delta(O_n)\} = \epsilon, \text{ independent of } \Theta,$$

then we say that $\delta(O_n)$ is a confidence region for Θ, with confidence coefficient ϵ.

It may be desired to estimate only a subset $\theta_1,\theta_2,\ldots,\theta_m$, $m < h$, of the h parameters (the remaining parameters are called nuisance parameters). Denote the m-dimensional space of $\Theta':(\theta_1,\theta_2,\ldots,\theta_m)$ by Ω'. If $\delta'(O_n)$ is a random region in Ω' such that

$$\Pr\{\Theta'\epsilon\delta'(O_n)\} = \epsilon, \text{ independent of } \Theta,$$

whatever be the true value Θ, then $\delta'(O_n)$ is said to be a confidence region for Θ' with confidence coefficient ϵ.

Example: Suppose O_{n_1} and O_{n_2} are samples from normal populations $N(a_1,\sigma^2)$ and $N(a_2,\sigma^2)$, respectively. We know from §5.25 that S_1/σ^2, S_2/σ^2 (defined in Example 2, §6.13),

$$\frac{n_1(\bar{x}_1-a_1)^2}{\sigma^2}, \quad \frac{n_2(\bar{x}_2-a_2)^2}{\sigma^2}$$

are independently distributed according to χ^2-laws with n_1-1, n_2-1, 1, 1 degrees of freedom respectively. By §5.23 it follows that

$$\frac{S_1 + S_2}{\sigma^2} \text{ and } \frac{n_1(\bar{x}_1-a_1)^2 + n_2(\bar{x}_2-a_2)^2}{\sigma^2}$$

are independently distributed according to χ^2-laws with $n_1 + n_2 - 2$ and 2 degrees of freedom respectively. Hence if we set

$$F = \frac{n_1(\bar{x}_1-a_1)^2 + n_2(\bar{x}_2-a_2)^2}{S_1 + S_2}\left(\frac{n-2}{2}\right),$$

then F is distributed according to $h'_{2,n-2}(F)$ where $n = n_1 + n_2$.

Therefore if F_ϵ is chosen so that

$$\int_0^{F_\epsilon} h_{2,n-2}(F)dF = \epsilon,$$

we may say that

$$\Pr\left(\frac{n_1(\bar{x}_1-a_1)^2+n_2(\bar{x}_2-a_2)^2}{S_1+S_2}\left(\frac{n-2}{2}\right) \leq F_\epsilon\right) = \epsilon,$$

which is equivalent to the statement that

$$\Pr\{(a_1,a_2)\in\delta(0_n)\} = \epsilon,$$

where $\delta(0_n)$ is the region in the (a_1,a_2) plane bounded by the random ellipse with equation

$$n_1(\bar{x}_1-a_1)^2 + n_2(\bar{x}_2-a_2)^2 = F_\epsilon \frac{2(S_1+S_2)}{n-2}.$$

In other words, the probability is ϵ that this ellipse will cover the true parameter point (a_1,a_2).

6.2 Point Estimation: Maximum Likelihood Statistics

Throughout this section we consider the point estimation of a parameter θ in the c. d. f. of a population. There may be other unknown parameters present, if so, we denote these by $\theta_2,\theta_3,\ldots,\theta_h$. A statistic is any function $T(0_n)$ of the sample, not depending on θ, or on any other parameters if such are present. Point estimation consists of the use of a single statistic for estimating the parameter; confidence intervals, we recall, involve two statistics, the end-points of the confidence interval, satisfying certain conditions (§6.1). Desirable conditions for statistics used as point estimates have been given by R. A. Fisher: An optimum estimate satisfies the criteria of consistency, efficiency, and sufficiency, defined below. A method which sometimes yields optimum statistics is Fisher's method of maximum likelihood.

6.21 Consistency

A statistic $T(0_n)$ is said to be a consistent estimate of θ if T converges stochastically (§4.21) to θ as $n \to \infty$. From Theorem (B) of §4.21 we know that whenever the population has finite variance, the sample mean is a consistent estimate of the population mean. We remark that consistency is purely an asymptotic property. If for every n, $E(T) = \theta$, then we say that the statistic T is unbiased. It follows from Theorem (A) of §4.21 that the sample mean is always an unbiased estimate of the population mean (whenever the latter exists). The following theorem enables us to recognize the consistency

of statistics in many cases:

Theorem (A): A sufficient condition that T be a consistent statistic for estimating θ is that $E(T) \to \theta$ and $\sigma_T^2 \to 0$ as $n \to \infty$.

To prove the theorem, write $T(O_n) = T_n$, and set any $\epsilon > 0$. We need to show that

$$\Pr(|T_n-\theta| > \epsilon) \to 0 \text{ as } n \to \infty.$$

Since $E(T_n) \to \theta$, there exists an N such that

$$|E(T_n)-\theta| < \tfrac{1}{2}\epsilon \text{ for } n > N.$$

We note that for $n > N$ the interval of T_n values $|T_n-E(T_n)| \leq \frac{1}{2}\epsilon$ is always contained in the interval $|T_n-\theta| \leq \epsilon$. Hence the probability of T_n falling outside the latter interval is $\leq$ the probability of T_n falling outside the former:

$$\Pr(|T_n-\theta| > \epsilon) \leq \Pr[|T_n-E(T_n)| > \tfrac{1}{2}\epsilon] = \Pr[|T_n-E(T_n)| > \sigma_T \cdot \epsilon/2\sigma_T].$$

By Tchebycheff's inequality (§2.71) the last expression is $\leq (2\sigma_T/\epsilon)^2$, and by hypothesis this $\to 0$ as $n \to \infty$ for all fixed $\epsilon > 0$.

Example: Let O_n be a sample from a population with an arbitrary c. d. f. about which we assume only that the fourth moment μ_4 about the mean exists, and consider s^2 as an estimate of σ^2, where

$$s^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1), \qquad \bar{x} = \sum_{i=1}^{n} x_i/n.$$

Then we know $E(s^2) = \sigma^2$, and by use of the well known formula for the variance of s^2,

$$\sigma^2_{s^2} = \tfrac{1}{n}[\mu_4 - \tfrac{n-3}{n-1}\sigma^4],$$

it follows from theorem (A) that s^2 is a consistent statistic for estimating σ^2. If we apply the same theorem to

$$(s')^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2/n,$$

we find it is also a consistent estimate of σ^2, but unlike s^2, it is biased.

6.22 Efficiency

$T(O_n)$ is said to be an efficient estimate of θ if

1) $\sqrt{n}(T-\theta)$ is asymptotically distributed according to $N(0,\mu)$ with $\mu < \infty$,

ii) for any other statistic $T'(0_n)$ such that $\sqrt{n}(T'-\theta)$ is asymptotically distributed according to $N(0,\mu')$, $\mu \leq \mu'$.

Since the asymptotic mean and variance of T are θ and μ/n, respectively, it follows from Theorem (A) of §6.21 that (i) implies the consistency of T. The efficiency of T' in estimating θ is defined by $E = \mu/\mu'$.

Example: Consider the sample mean $\bar{x}$ and the sample median $\tilde{x}$ of 0_n from $N(a,\sigma^2)$ as estimates of a. We have from §5.11 that $\sqrt{n}(\bar{x}-a)$ is distributed according to $N(0,\sigma^2)$, and from §4.53 that $\sqrt{n}(\tilde{x}-a)$ is asymptotically distributed according to $N(0,\frac{1}{2}\pi\sigma^2)$. Hence $\bar{x}$ is more efficient than $\tilde{x}$. However, to prove $\bar{x}$ "efficient" it would be necessary to verify condition (ii) of the definition. This example may be generalized as follows: If 0_n is from a population with p. d. f. $f(x)$, if the population median = a (the population mean), and if $f(x)$ is continuous at a, then using the results of §4.53 on the asymptotic distribution of $\tilde{x}$, we find that $\bar{x}$ is a more efficient estimate of a than $\tilde{x}$ if $\sigma < [2f(a)]^{-1}$, $\tilde{x}$ is more efficient if $\sigma > [2f(a)]^{-1}$.

6.23 Sufficiency

T is said to be a sufficient statistic for estimating θ if for any other statistic T', the conditional distribution $f(T'|T)$ of T', given T, is independent of θ. (We use the same notation $f(T'|T)$ whether the population is continuous or discrete.) Thus, expected values, moments and other probability calculations about T', given T, will be calculated from $f(T'|T)$ and hence will not depend on θ, but they will depend on T in general. Or, in Fisher's terminology a sufficient statistic "exhausts the information" in a sample. We note that sufficiency, unlike consistency and efficiency, is not merely an asymptotic property.

A convenient method of spotting sufficient statistics is embodied in

Theorem (A)*: If the population distribution is continuous, let $P(0_n;\theta,\theta_2,..,\theta_h)$ be the p. d. f. of 0_n; if the distribution is discrete, let $P(0_n;\theta,\theta_2,\ldots,\theta_h)$ be the discrete probability of 0_n. In either case a necessary and sufficient condition that T be a sufficient statistic for estimating θ is that the function P factor in the following manner

$$P(0_n;\theta,\theta_2,\ldots,\theta_h) = g_1(T;\theta,\theta_2,\ldots,\theta_h)\cdot g_2(0_n;\theta_2,\theta_3,\ldots,\theta_h).$$

A sufficient set of statistics with regard to a set of parameters may be defined and an analogue of Theorem(A)obtained for that case; see Neyman and Pearson, Statistical Research Memoirs, Vol. 1 (1936), pp. 119-121.

---

*For proof, see J. Neyman, Giornale dell Istituto Italiano degli Attuari, Vol. 6 (1934), pp. 320-334.

Example 1: Suppose O_n is from $N(a,\sigma^2)$. Then

$$P(O_n;a,\sigma^2) = e^{-\frac{n}{2}(\bar{x}-a)^2/\sigma^2} \cdot (2\pi\sigma^2)^{-\frac{1}{2}n} e^{-\frac{1}{2}S/\sigma^2},$$

where

$$S = \sum_{i=1}^{n}(x_i-\bar{x})^2.$$

(Here and in the following examples the factors corresponding to g_1 and g_2 of Theorem (A) are separated by a dot.) Hence $\bar{x}$ is a sufficient statistic for estimating a. In this case there is no sufficient statistic for σ^2 but it is easily shown that $\bar{x}, S/(n-1)$ are a sufficient set of (unbiased) statistics for a and σ^2.

Example 2: For O_n from $N(0,\theta)$,

$$P(O_n;\theta) = (2\pi\theta)^{-\frac{1}{2}n} e^{-\frac{1}{2}S'/\theta} \cdot 1,$$

where $S' = \sum_{i=1}^{n} x_i^2$. Hence S' is a sufficient statistic for estimating θ. S'/n is an unbiased (see ex.2, §6.24) sufficient statistic.

Example 3: Suppose the population has the discrete distribution

$$p(x;\theta) = \theta^x e^{-\theta}/x!, \qquad x = 0,1,2,\ldots$$

We recall from §3.13 that $E(x) = \theta$. For a sample $O_n:(x_1,x_2,\ldots,x_n)$,

$$P(O_n;\theta) = \prod_{i=1}^{n} \theta^{x_i} e^{-\theta}/x_i! .$$

If we write this

$$P(O_n;\theta) = (\theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}) \cdot (1/\prod_{i=1}^{n} x_i!),$$

we see that $\sum_{i=1}^{n} x_i$ is a sufficient statistic for estimating θ. Since

$$E(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} E(x_i) = n\theta,$$

it follows that $\bar{x} = \sum_{i=1}^{n} x_i/n$ is an unbiased sufficient statistic.

6.24 Maximum Likelihood Estimates

The function $P(O_n;\theta,\theta_2,\ldots,\theta_h)$ defined in Theorem (A) of §6.23, when considered as a function of the parameter point $\theta,\theta_2,\ldots,\theta_h$, for fixed O_n, is called the likelihood of the parameter point. If the likelihood function P has a unique maximum at $\theta = \hat{\theta}(O_n)$, $\theta_2 = \hat{\theta}_2(O_n),\ldots,\theta_h = \hat{\theta}_h(O_n)$, then the set of statistics $\theta,\theta_2,\ldots,\theta_h$ is called the maximum likelihood estimate of the parameter point.

Let us consider the case of one parameter , say θ. In Theorem (A), §6.12, it was shown that under certain conditions the quantity $\frac{1}{\sqrt{n}A}\frac{\partial \log P}{\partial \theta}$ is asymptotically distributed according to N(0,1). Let us assume that $\hat{\theta}$ is the value of θ which maximizes P and that we can make the following expansion about $\theta = \hat{\theta}$,

$$(a)\qquad \frac{1}{\sqrt{n}A}\left(\frac{\partial \log P}{\partial \theta}\right) = \frac{1}{\sqrt{n}A}\left(\frac{\partial \log P}{\partial \theta}\right)_{\hat{\theta}} + \frac{1}{\sqrt{n}A}\left(\frac{\partial^2 \log P}{\partial \theta^2}\right)_{\hat{\theta}}(\theta-\hat{\theta}) + \frac{1}{2\sqrt{n}A}\left(\frac{\partial^3 \log P}{\partial \theta^3}\right)_{\tilde{\theta}}(\theta-\hat{\theta})^2$$

$$= -AV + UV + \frac{1}{\sqrt{n}}WV^2,$$

where $\tilde{\theta}$ is on the interval $(\theta,\hat{\theta})$, and

$$U = \frac{1}{A}\left[A^2+\left(\frac{1}{n}\frac{\partial^2 \log P}{\partial \theta^2}\right)_{\hat{\theta}}\right],$$

$$V = \sqrt{n}(\theta-\hat{\theta}), \qquad W = \frac{1}{2A}\left(\frac{1}{n}\frac{\partial^3 \log P}{\partial \theta^3}\right)_{\tilde{\theta}}.$$

We have employed the fact that $\partial P/\partial\theta$ vanishes for $\theta = \hat{\theta}$. Now from Theorem (A), §6.12,

$$(b)\qquad \Pr\left(\frac{1}{\sqrt{n}A}\frac{\partial \log P}{\partial \theta} < d\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{d} e^{-\frac{1}{2}x^2}\,dx + \eta_n,$$

where $\eta_n \longrightarrow 0$ as $n \longrightarrow \infty$.

Making use of (a) we may write

$$(c)\qquad \Pr\left(-AV + UV + \frac{1}{\sqrt{n}}WV^2 < d\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{d} e^{-\frac{1}{2}x^2}\,dx + \eta_n.$$

Considering U, V, W as three random variables, the left side of (c) states that the probability of U, V, W falling into a certain region in this space is given by the expression on the right. Now let us assume (1) that $\left(\frac{1}{n}\frac{\partial^2 \log P}{\partial \theta^2}\right)_{\hat{\theta}}$ converges stochastically to $E\left(\frac{\partial^2 \log f(x,\theta)}{\partial \theta^2}\right)$ which we shall assume $= -A^2$ (implying that U converges stochastically to 0) as $n \longrightarrow \infty$, (2) that $\left(\frac{1}{n}\frac{\partial^3 \log P}{\partial \theta^3}\right)_{\tilde{\theta}}$ (and hence 2AW) converges stochastically to some finite number K, and (3) that V has some limiting non-degenerate p. d. f. as $n \longrightarrow \infty$ (i. e., has a c. d. f. which is continuous), then the limiting form of the distribution function in the U, V, W space as $n \longrightarrow \infty$ is a one-dimensional p. d. f. along the straight line

$$\begin{cases} U = 0 \\ W = K \end{cases}.$$

The p. d. f. on this line is that of the limiting distribution of V. Hence,

$$\lim_{n \to \infty} \Pr\left(-AV + UV + \frac{1}{\sqrt{n}} WV^2 < d \right) = \lim_{n \to \infty} \Pr(-AV < d).$$

The equality of the two expressions for A^2 is a reasonable assumption as the reader will see from the following discussion:

(d) $$f \frac{\partial \log f}{\partial \theta} = \frac{\partial f}{\partial \theta} .$$

Differentiating this with respect to θ, we get

(e) $$f \frac{\partial^2 \log f}{\partial \theta^2} + \frac{\partial f}{\partial \theta} \frac{\partial \log f}{\partial \theta} = \frac{\partial^2 f}{\partial \theta^2} .$$

Substituting (d) into (e), and integrating with respect to x from $-\infty$ to $+\infty$, we have

$$E\left[\frac{\partial^2 \log f}{\partial \theta^2}\right] + E\left[\left(\frac{\partial \log f}{\partial \theta}\right)^2\right] = \int_{-\infty}^{+\infty} \frac{\partial^2 f}{\partial \theta^2} \, dx.$$

Now if we may interchange the order of integration and differentiation in the right member, then the left member is seen to be equal to

$$\frac{\partial^2}{\partial \theta^2} \int_{-\infty}^{\infty} f \, dx = \frac{\partial^2}{\partial \theta^2} (1) = 0.$$

We may summarize in the following

Theorem (A): Let $O_n(x_1, x_2, \ldots, x_n)$ be a sample from a population with c. d. f. $F(x;\theta)$. Let $P(O_n) = \prod_{i=1}^{n} f(x_i;\theta)$ be the likelihood function, where $f(x;\theta)$ is the p. d. f. if x is a continuous variable and probability of x if x is discrete. Let $P(O_n;\theta)$ have a unique maximum at $\theta = \hat{\theta}$, and assume

(i) $$E\left[\left(\frac{\partial \log f}{\partial \theta}\right)^2\right] = -E\left(\frac{\partial^2 \log f}{\partial \theta^2}\right) = A^2,$$

and that as $n \to \infty$

(ii) $\left(\frac{1}{n} \frac{\partial^2 \log P}{\partial \theta^2}\right)_{\hat{\theta}}$ converges stochastically to $-A^2$,

(iii) $\left(\frac{1}{n} \frac{\partial^3 \log P}{\partial \theta^3}\right)_{\theta}$ converges stochastically to a finite K,

(iv) $\sqrt{n}(\hat{\theta} - \theta)$ has a limiting non-degenerate p. d. f.

Then $\sqrt{n}(\hat{\theta}-\theta)$ is distributed asymptotically according to $N(0,\frac{1}{A^2})$.

Under fairly general conditions, which will not be given here, it can be shown that if $\bar{\theta}$ is any other statistic such that $\sqrt{n}(\bar{\theta}-\theta)$ is asymptotically distributed according to $N(0,B^2)$, then $B^2 \geq \frac{1}{A^2}$.

In the present case where the c. d. f. $F(x;\theta)$ depends on only one parameter it is often possible to transform from the old parameter θ to a new parameter ϕ so that the asymptotic variance of $\hat{\phi}$, the maximum likelihood estimate of ϕ will be independent of the parameter. Let A^2 be defined as before, let $\phi = h(\theta)$, a function to be determined, and define

$$B^2 = E[(\frac{\partial \log f}{\partial \phi})^2].$$

We will try to determine the function $h(\theta)$ so that B^2 is a given positive constant. We have

$$A^2 = E[(\frac{\partial \log f}{\partial \theta})^2] = E[(\frac{\partial \log f}{\partial \phi} \cdot \frac{d\phi}{d\theta})^2] = B^2(\frac{d\phi}{d\theta})^2,$$

$$\frac{d\phi}{d\theta} = \frac{A}{B},$$

$$\phi = \frac{1}{B}\int A d\theta + C,$$

where C is any arbitrary constant. If the last equation determines ϕ as a monotonic continuous function of θ, then since $P(O_n;\theta)$ has a unique maximum for $\theta = \hat{\theta}$, clearly $P(O_n;h^{-1}(\phi))$ has a unique maximum for $\phi = \hat{\phi} = h(\hat{\theta})$. By Theorem (A) the asymptotic variance of $\hat{\phi}$, the maximum likelihood estimate of ϕ, will be $(nB^2)^{-1}$ which is independent of ϕ. As an illustration the reader can verify that in Example 2 below, $\phi = \log \theta$ is a new parameter of the desired type.

Theorem (A), §6.12, and Theorem (A) of the present section can both be extended to the case of several parameters. In the case of several parameters it may be shown under conditions analogous to those in Theorem (A) that for large n $\sqrt{n}(\hat{\theta}_1-\theta_1)$, $\sqrt{n}(\hat{\theta}_2-\theta_2),\ldots,$ $\sqrt{n}(\hat{\theta}_h-\theta_h)$ are asymptotically distributed according to a normal multivariate distribution with variance-covariance matrix $||\sigma_i\sigma_j\rho_{ij}||$ given by $||A_{ij}||^{-1}$, where

$$(f) \qquad A_{ij} = E(\frac{\partial \log f}{\partial \theta_i} \cdot \frac{\partial \log f}{\partial \theta_j}) = -E(\frac{\partial^2 \log f}{\partial \theta_i \partial \theta_j}).$$

Example 1: Suppose O_n is from $N(a,1)$:

$$f(x;a) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-a)^2},$$

$$\log f = -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}(x-a)^2,$$

$$\partial \log f/\partial a = x - a.$$

If we use the first of the two expressions for A^2 we get

$$A^2 = E[(x-a)^2] = \sigma^2 = 1.$$

To use the second expression we would have to take the expected value of

$$-\partial^2 \log f/\partial a^2 = 1,$$

and we note we get the same result. To find $\hat{\theta}$ we inspect

$$P(0_n;a) = (2\pi)^{-\frac{1}{2}n} e^{-\frac{1}{2}[n(\bar{x}-a)^2+S]},$$

and see that this is maximum when the exponent is minimum, that is for

$$\hat{a} = \bar{x}.$$

Theorem (A) says that $\sqrt{n}(\bar{x}-a)$ is asymptotically distributed according to $N(0,1)$. In the present case this is the <u>exact</u> distribution.

<u>Example 2</u>: For 0_n from $N(0,\theta)$,

$$f(x;\theta) = (2\pi\theta)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2/\theta},$$

$$\log f = -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log\theta - \tfrac{1}{2}x^2/\theta,$$

$$\partial \log f/\partial\theta = \tfrac{1}{2}(-1/\theta + x^2/\theta^2),$$

$$A^2 = \tfrac{1}{4}E[(-1/\theta + x^2/\theta^2)^2].$$

Let us see whether it may not be easier to calculate A^2 from the other formula:

$$\partial^2 \log f/\partial\theta^2 = \tfrac{1}{2}\theta^{-2} - x^2/\theta^3,$$

$$A^2 = -E(\tfrac{1}{2}\theta^{-2} - x^2/\theta^3) = -\tfrac{1}{2}\theta^{-2} + E(x^2/\theta)/\theta^2.$$

Since x^2/θ has the χ^2-distribution with $k = 1$ degrees of freedom, its mean is $k = 1$. Hence

$$A^2 = -\tfrac{1}{2}\theta^{-2} + \theta^{-2} = \tfrac{1}{2}\theta^{-2}.$$

Now

$$P(0_n;\theta) = (2\pi)^{-\frac{1}{2}n}\theta^{-\frac{1}{2}n}e^{-\frac{1}{2}S'/\theta},$$

where $S' = \sum_{i=1}^{n} x_i^2$. Differentiating

$$\log P = -\frac{n}{2}\log(2\pi) - \frac{1}{2}n\log\theta - \frac{1}{2}S'/\theta$$

with respect to θ, we get

$$\partial P/\partial\theta = \tfrac{1}{2}P\cdot(-n/\theta+S'/\theta^2).$$

Equating this to zero and solving for θ, we find

$$\hat{\theta} = S'/n.$$

By Theorem (A), $\sqrt{n}(\hat{\theta}-\theta)$ is asymptotically distributed according to $N(0,2\theta^2)$. Since S'/θ actually has the χ^2-distribution with n degrees of freedom, its exact mean and variance are n and 2n, respectively; hence the asymptotic mean and variance given by Theorem (A) turn out to be the exact mean and variance. However, the exact distribution of $n\hat{\theta}/\theta$ is the χ^2-distribution with n degrees of freedom, and not a normal distribution.

Example 3: As an illustration of the method of obtaining maximum likelihood estimates when the distribution is discrete, consider again the sample of Example 3, §6.23. We may write

$$P(0_n;\theta) = \theta^{n\bar{x}}\,\theta^{-n\theta}\,U,$$

where

$$\bar{x} = \sum_{i=1}^{n} x_i/n,$$

$$U = 1/\prod_{i=1}^{n} x_i!$$

are independent of θ. To find $\hat{\theta}$ we set $\partial P/\partial\theta = 0$ and solve for θ:

$$\log P = n\bar{x}\log\theta - n\theta + \log U,$$

$$\partial P/\partial\theta = P\cdot(n\bar{x}/\theta - n) = 0,$$

$$\hat{\theta} = \bar{x}.$$

This we have already shown to be an unbiased sufficient statistic. We calculate

$$\log f(x,\theta) = x\log\theta - \theta - \log x!$$

$$\partial^2\log f/\partial\theta^2 = -x/\theta^2,$$

$$A^2 = -E(-x/\theta^2) = 1/\theta.$$

Thus Theorem (A) tells us that $\sqrt{n}(\bar{x}-\theta)$ is asymptotically distributed according to $N(0,\theta)$.

Example 4: In this example we illustrate the method of maximum likelihood for obtaining estimates when more than one parameter is present in the population distribution. Suppose O_n is from $N(a,\theta)$. Then

$$P(O_n;a,\theta) = (2\pi\theta)^{-\frac{1}{2}n} e^{-\frac{1}{2}[n(\bar{x}-a)^2+S]/\theta},$$

where

$$S = \sum_{i=1}^{n}(x_i-\bar{x})^2.$$

To find the estimates $\hat{a}$, $\hat{\theta}$, we set

$$\partial P/\partial a = \partial P/\partial\theta = 0$$

and solve for a and θ:

$$\log P = -\tfrac{1}{2}n\log(2\pi) - \tfrac{1}{2}n\log\theta - \tfrac{1}{2}[n(\bar{x}-a)^2+S]/\theta,$$

$$\partial P/\partial a = P[n(\bar{x}-a)/\theta] = 0,$$

$$\partial P/\partial\theta = \tfrac{1}{2}P\{-n/\theta + [n(\bar{x}-a)^2+S]/\theta^2\} = 0.$$

The solutions of these equations are easily found to be

$$\hat{a} = \bar{x}, \qquad \hat{\theta} = S/n.$$

As we have previously noted, these are both consistent estimates, but the latter is biased.

Let us compute the asymptotic variance-covariance matrix of $\sqrt{n}(\hat{a}-a)$ and $\sqrt{n}(\hat{\theta}-\theta)$ as given in the generalization stated below Theorem (A):

$$\log f = -\tfrac{1}{2}\log\theta - \tfrac{1}{2}(x-a)^2/\theta - \tfrac{1}{2}\log(2\pi),$$

$$\frac{\partial^2\log f}{\partial a^2} = -\frac{1}{\theta}, \quad \frac{\partial^2\log f}{\partial\theta^2} = \frac{1}{2\theta^2} - \frac{(x-a)^2}{\theta^3}, \quad \frac{\partial^2\log f}{\partial a\partial\theta} = -\frac{x-a}{\theta^2}.$$

Hence the asymptotic variance-covariance matrix is

$$\left\| \begin{matrix} A_{aa} & A_{a\theta} \\ A_{\theta a} & A_{\theta\theta} \end{matrix} \right\|^{-1} = \left\| \begin{matrix} \frac{1}{\theta} & 0 \\ 0 & \frac{1}{2\theta^2} \end{matrix} \right\|^{-1} = \left\| \begin{matrix} \theta & 0 \\ 0 & 2\theta^2 \end{matrix} \right\|$$

It is easily verified that the entries in the last matrix are exact with the exception of the (2,2) entry whose exact value is $2\theta^2(n-1)/n$.

6.3 Tolerance Interval Estimation

In the foregoing sections we have discussed two methods of estimating one or more parameters in distribution functions from samples: the method of confidence intervals and the method of point estimation based on the method of maximum likelihood. If

the original parameters, say $\theta_1, \theta_2, \ldots, \theta_h$, are transformed to new parameters $\phi_1, \phi_2, \ldots, \phi_h$ by any one-to-one transformation $\phi_i = \phi_i(\theta_1, \theta_2, \ldots, \theta_h)$, $i = 1,2,\ldots,h$, which is continuous and possesses first derivatives, we may apply both methods of estimation as before to the problem of estimating the new parameters. In fact, it can be readily verified that the maximum likelihood estimates of the ϕ_i are $\phi_i(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_h)$, $i = 1,2,\ldots,h$, where $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_h$ are maximum likelihood estimates of the θ_i. A specific case of transforming a single parameter was discussed in §6.24; the problem there was to find a function of θ having a maximum likelihood estimate whose variance in large samples (to terms of order $\frac{1}{n}$) does not depend on this function of θ.

Another problem of estimating a function of the parameter which deserves special comment is that of setting tolerance limits (see §4.55). This problem is as follows: Suppose $f(x,\theta)dx$ is the probability element of x where θ is the parameter. For a given $0 < \beta' < 1$ let L_1 and L_2 be such that

$$\int_{-\infty}^{L_1} f(x,\theta)dx = \frac{1-\beta'}{2}, \qquad \int_{L_2}^{\infty} f(x,\theta)dx = \frac{1-\beta'}{2}.$$

L_1 and L_2 are continuous functions of β' and θ; denote them by $L_1(\theta,\beta')$, $L_2(\theta,\beta')$. From the discussion in the paragraph above it follows that in a sample of size n the likelihood estimates of $L_1(\theta,\beta')$ and of $L_2(\theta,\beta')$ are $L_1(\hat{\theta},\beta')$ and $L_2(\hat{\theta},\beta')$, which are completely expressible in terms of the sample, when the functional form of $f(x,\theta)$ is given. Now the integral

$$\text{(a)} \qquad v = \int_{L_1(\hat{\theta},\beta')}^{L_2(\hat{\theta},\beta')} f(x,\theta)dx$$

is a random variable which represents the proportion of the population for which $L_1(\hat{\theta},\beta') < x < L_2(\hat{\theta},\beta')$. Assume that the distribution function of the integral (a) is independent of θ. If $\hat{\theta}$ converges stochastically to θ as $n \to \infty$ (which is implied by the assumption (iv), Theorem (A), §6.24) then $L_1(\hat{\theta},\beta')$ and $L_2(\hat{\theta},\beta')$ converges stochastically to $L_1(\theta,\beta')$ and $L_2(\theta,\beta')$, respectively, and hence the integral (a) converges stochastically to β'. Therefore, for a given β on the interval $(0,1)$, and ϵ on the same interval and choosing some β' on the interval $(\beta,1)$, one can choose an n, say n', such that for $n > n'$

$$\text{(b)} \qquad \Pr(v \geq \beta) > \epsilon,$$

no matter what value θ may have. For some values of β and ϵ, particularly those near unity (e. g., .95 or .99) there exists a smallest n, say n_0, such that

(c) $$\Pr(v \geq \beta) \geq \epsilon.$$

Therefore, under this condition $L_1(\hat{\theta},\beta')$ and $L_2(\hat{\theta},\beta')$ are 100β % parameter-free tolerance limits at probability level ϵ (see §4.55)*. The interval $L_1(\hat{\theta},\beta')$, $L_2(\hat{\theta},\beta')$ on the x-axis may be referred to as a tolerance interval based on samples of size n_o for covering or estimating at least 100β % of the values of x of the population, with a probability of at least ϵ. These results may be extended to the case in which two or more parameters are involved in the distribution function of x.

It is evident that there are many ways of choosing tolerance limits as functions of θ so that statement (b) can be made, e. g., L_1 and L_2 could be determined by cutting off unequal probabilities from the tail of the distribution function $f(x,\theta)$ rather than equal probabilities.

The reader should note carefully the distinction between a confidence interval statement (§6.11) about a population parameter and a tolerance interval statement (in this § and §4.55) about a population proportion. It will be seen, however, in the last example of §6.12 that the confidence statement about the proportion p in a binomial population is closely analogous to a tolerance interval statement about a population proportion in the case of a population with a continuous random variable.

As an example of tolerance limits of this type involving two parameters consider a sample $O_n(x_1, x_2, \ldots, x_n)$ drawn from the normal distribution $N(a,\sigma^2)$. Let $\bar{x}$ be the sample mean and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Let $t_{\beta'}$ be such that

(d) $$\int_{-t_{\beta'}}^{t_{\beta'}} g_{n-1}(t)dt = \beta',$$

where $g_{n-1}(t)$ is the "Student" distribution with n-1 degrees of freedom (see §5.3). Let $L_1 = \bar{x} - t_{\beta'}\sqrt{\frac{n+1}{n}}\,s$, and $L_2 = \bar{x} + t_{\beta'}\sqrt{\frac{n+1}{n}}\,s$. The proportion of the population having values of x on the tolerance interval $\bar{x} \pm t_{\beta'}\sqrt{\frac{n+1}{n}}\,s$ is

(e) $$\int_{L_1}^{L_2} N(a,\sigma^2)dx.$$

The distribution function of this integral is not known. However, it has been shown

*For details of the approach to tolerance limits for large samples when the functional form of $f(x,\theta)$ is known, see A. Wald, "Setting of Tolerance Limits when the Sample is Large", Annals of Math. Stat., Vol. 13 (1942).

**S. S. Wilks, "Determination of Sample Size for Setting Tolerance Limits", Annals of Math. Stat., Vol. 12, (1941), pp. 94-95.

that the mean value of this integral is β. Its variance has been determined only for large samples, which is $t_\beta^2 e^{-t_\beta^2}/(\pi n)$ to terms of order $\frac{1}{n}$.

In the discussion thus far, it has been assumed that the functional form of the population distribution $f(x,\theta)$ is known but the value of θ is unknown. From the point of view of practical statistics the case in which x is a continuous random variable with an unknown distribution is perhaps more important than the case in which the functional form is known. This case has been treated in §4.55*.

6.4 The Fitting of Distribution Functions

The problem of fitting of distribution functions is as follows: Let $F(x,\theta_1,\theta_2,\ldots,\theta_h)$ be a c. d. f. depending on the h parameters $\theta_1, \theta_2,\ldots, \theta_h$, and let O_n be a sample of size n from a population having this c. d. f. Consider the values $x_1, x_2,\ldots, x_n$ of the sample O_n as n values of a variable x. From these n values we can construct an "empirical" c. d. f., say $F_n(x)$. The problem of fitting $F(x,\theta_1,\ldots,\theta_h)$ to $F_n(x)$ is that of determining $\theta_1, \theta_2,\ldots,\theta_h$ so that $F(x,\theta_1,\ldots,\theta_h)$ is approximately equal to $F_n(x)$ in some sense.

The method of maximum likelihood provides one method of determining values of $\theta_1, \theta_2,\ldots, \theta_h$ by maximizing the likelihood $\prod_{i=1}^{n} f(x_i,\theta_1,\theta_2,\ldots,\theta_h)$ with respect to the θ's. Clearly the values assigned to the parameters by this method are precisely their maximum likelihood estimates $\hat{\theta}_1, \hat{\theta}_2,\ldots, \hat{\theta}_h$ (§6.24). This method of fitting is best in the sense that for large n, the variance of each θ_i is less than or equal to that of any other consistent and asymptotically normally distributed estimate of θ_i.

Another method of fitting which is easy to apply in many problems is the method of moments. This method consists of equating the moments

$$\mu_i' = M_i' \qquad (i = 1,2,\ldots,h)$$

and solving for $\theta_1, \theta_2,\ldots, \theta_h$ (assuming μ_i' exists for $i = 1,2,\ldots,h$), where

*For further details, not given here, the reader is referred to S. S. Wilks, loc. cit., and also S. S. Wilks, "Statistical Prediction with Special Reference to the Problem of Tolerance Limits", Annals of Math. Stat., Vol. 13 (1942). An extension of the notion of tolerance limits to two or more variables is to be presented in a forthcoming paper in the Annals of Math. Stat. by A. Wald.

$$\mu_i' = \int_{-\infty}^{\infty} x^i dF(x,\theta_1,\theta_2,\ldots,\theta_h),$$

$$M_i' = \int_{-\infty}^{\infty} x^i dF_n(x) = \sum_{\alpha=1}^{n} x_\alpha^i/n.$$

In the case of fitting certain distributions, for example the normal distribution, the binomial and Poisson distributions, the two methods of fitting yield the same results.

CHAPTER VII

TESTS OF STATISTICAL HYPOTHESES

Suppose the distribution function of a population depends on parameters θ_1, θ_2, ..., θ_h. We assume the functional form of the distribution to be known, but not the true parameter values. Let Ω be the h-dimensional space of admissible parameter values. Denote the parameter point by Θ. Let ω be a specified point set in Ω: it may be of dimensionality 0,1,..., up to h. In this chapter we consider tests of the statistical hypothesis,

$$H_0: \Theta \in \omega.$$

A test of H_0 is a procedure for accepting or rejecting H_0 on the evidence afforded by a sample from the population. A more precise definition of a test will be given in §7.3. As a general rule one sets up a test with the hope of rejecting the hypothesis, and for this reason the hypothesis is often called a null hypothesis in such cases. Thus, if one desires confirmation of a suspicion that two populations have different means, one takes as H_0 the hypothesis that the means are equal, and if H_0 is rejected by the test, then one's suspicion is confirmed on the basis of the test used.

Statistical hypotheses are classified as follows: If ω is a single point of Ω, that is, if H_0 states $\Theta = \Theta_0$, then H_0 is called simple; in any other case H_0 is called composite.

7.1 Statistical Tests Related to Confidence Intervals

Consider the case where H_0 specifies the value of only one parameter θ_1,

$$H_0: \quad \theta_1 = \theta_1^0.$$

If the population distribution depends on no other parameters, this is a simple hypothesis; if other parameters $\theta_2, \ldots, \theta_h$ are present, H_0 is composite, ω being the h-1 dimensional

subspace (hyperplane) in Ω defined by $\theta_1 = \theta_1^0$. If confidence intervals $\delta(O_n)$ for θ_1 are available, then one may proceed as follows: Form $\delta(O_n)$ for the sample O_n, and reject H_0 unless $\delta(O_n)$ covers θ_1^0. If ϵ is the confidence coefficient, then

$$\Pr(\text{rejecting } H_0 \text{ if it is true}) = 1 - \Pr\{\theta_1^0 \epsilon \delta(O_n) | \theta_1 = \theta_1^0\} = 1 - \epsilon.$$

The quantity $\alpha = 1 - \epsilon$ is called the <u>significance level</u> of the test. It will be noted that when confidence intervals for θ_1^0 are known, then a whole family of tests is at hand: A test exists for every θ_1^0, that is, for every admissible value of θ_1. We remark that beyond the statement $\Pr(\text{rejecting } H_0 \text{ if true}) = \alpha$, no further property of the test can be deduced from the definition of confidence intervals. One might ask about the Pr(accepting H_0 if false), that is, accepting H_0 when θ_1 has some other value than θ_1^0, but the significance level tells us nothing about this*. As will be seen in the examples below, our method usually leads us to the calculation of a certain statistic, say T, and H_0 is rejected if T falls in a certain range R. Suppose, for example, that R is the range $T > T_0$, and that T possesses the p. d. f. f(T) if $\theta_1 = \theta_1^0$. In certain cases it is sometimes said that $\alpha = \Pr(\text{finding a value of T less probable than } T_0 \text{ if } H_0 \text{ is true})$. This really does not motivate the test any better: If by "T_1 is less probable than T_2" we mean $f(T_1) < f(T_2)$, then the same test can be made with other statistics $S = \phi(T)$, and the relation "less probable" is not invariant under such transformations**.

It should be noted that confidence intervals give us a far more complete judgement about the parameter θ_1 than significance tests. We also remark that if confidence regions (§6.14) for the set $\theta_1, \theta_2, \ldots, \theta_m$ are available, then so are significance tests for the hypothesis

$$H_0: \quad \theta_1 = \theta_1^0, \quad \theta_2 = \theta_2^0, \ldots, \theta_m = \theta_m^0.$$

H_0 is simple if $m = h$, composite if $m < h$.

<u>Example 1</u>: Suppose that on the basis of the sample O_n from a population with the distribution $N(a,\sigma^2)$, where a and σ^2 are unknown, we wish to test the (Student) hypothesis

$$H_0: \quad a = a_0.$$

This is a composite hypothesis: The space Ω of admissible parameter points (a,σ^2) is

---

*See §7.3.

**This may be shown by considering the signs of f'(T) and g'(S), where g(S) is the p. d. f. of S.

(b) $$P_o = P(O_n;a_o,\theta) = (2\pi\theta)^{-\frac{1}{2}n} e^{-\frac{1}{2}[n(\bar{x}-a_o)^2+S]/\theta},$$

$$\log P_o = -\frac{1}{2} n \log (2\pi) - \frac{1}{2}n \log \theta - \frac{1}{2}[n(\bar{x}-a_o)^2+S]/\theta,$$

$$\partial P_o/\partial\theta = P_o\{-\frac{1}{2}n/\theta + \frac{1}{2}[n(\bar{x}-a_o)^2+S]/\theta^2\}.$$

Equating this to zero and solving for θ, we get

$$\theta = (\bar{x}-a_o)^2+S/n,$$

and substituting this into (b), we find

$$P_\omega(O_n) = \{2\pi[(\bar{x}-a_o)^2+S/n]\}^{-\frac{1}{2}n} e^{-\frac{1}{2}n}.$$

Hence

$$\lambda = [1+n(\bar{x}-a_o)^2/S]^{-\frac{1}{2}n}.$$

The distribution of λ under the assumption that H_o is true is independent of the unknown θ, in fact

$$\lambda = [1+t^2/(n-1)]^{-\frac{1}{2}n},$$

where

$$t = \sqrt{n}(\bar{x}-a_o)/s$$

has the t-distribution $g_{n-1}(t)$ with n-1 degrees of freedom. Let $t^2 = t_o^2$ correspond to $\lambda = \lambda_o$. Then $\lambda \leqq \lambda_o$ if and only if $|t| > t_o$. To get

$$Pr(\lambda \leqq \lambda_o) = \alpha,$$

we define t_o from

$$2\int_{t_o}^{\infty} g_{n-1}(t)dt = \alpha.$$

The likelihood ratio test for H_o is seen to be the same as the (Student) test of Example 1, §7.1.

In many cases the asymptotic distribution of the likelihood ratio is given by

Theorem (A): Suppose the c. d. f. of the population depends on parameters θ_1, $\theta_2,\ldots,\theta_h$, and that λ is the likelihood ratio for the hypothesis

$$H_o: \theta_1 = \theta_1^o, \theta_2 = \theta_2^o,\ldots, \theta_m = \theta_m^o,$$

where $m \leq h$. Then under certain regularity conditions* the asymptotic distribution of $-2 \log \lambda$, under the assumption that H_0 is true, is the χ^2-distribution with m degrees of freedom.

In the above example we may write

$$\lambda = (1 + \tfrac{1}{2}t^2/N)^{-N} (1 + \tfrac{1}{2}t^2/N)^{-\frac{1}{2}},$$

where $N = \frac{1}{2}(n-1)$. Hence as $n \longrightarrow \infty$, $N \longrightarrow \infty$, and

$$\lambda \sim e^{-\frac{1}{2}t^2},$$

$$-2 \log \lambda \sim t^2.$$

Since the asymptotic distribution of t is N(0,1), the asymptotic distribution of t^2 is the χ^2-distribution with one degree of freedom, and this accords with theorem (A).

7.3 The Neyman-Pearson Theory of Testing Hypotheses

In the notation introduced at the beginning of Chapter VII, consider the hypothesis

$$H_0: \Theta \in \omega.$$

In many problems (for instance, all the examples we have considered in §7.1) several tests, or a whole family of tests, are available, and the question arises, which is the "best" test? For the comparison of tests, Neyman and Pearson have introduced the concept of the power of a test. We approach this concept through the following steps:

First, we note that any test consists of the choice of a (B-meas.) region w in the sample space and the rule that we reject H_0 if and only if the sample point O_n falls in w. w is called the critical region of the test. The power of the test is defined to be the probability that we reject H_0. This is a function of the critical region w (a set function of w) and of the parameter point Θ (a point function of Θ). We write it $P(w|\Theta)$ and note it is

$$P(w|\Theta) = \Pr(O_n \in w|\Theta).$$

The interpretation of the power function is based on the following observation: In using a test of H_0, two types of error are possible (exhaustive and mutually exclusive): (I) We may reject H_0 when it is true. (II) We may accept H_0 when it is false, i. e., when Θ is a point not in ω. We call these respectively Type I and Type II errors. Now a

*The regularity conditions are the same as those for the multi-parameter analogue of Theorem (A), §6.24.

Type I error can only occur if the true $\Theta \in \omega$. Hence the probability of making a Type I error if $\Theta \in \omega$ is

(a) $$\Pr(O_n \in w | \Theta \in \omega) = P(w|\Theta) \text{ for } \Theta \in \omega.$$

A Type II error can be committed only if $\Theta \notin \omega$. The probability of making a Type II error if $\Theta \notin \omega$ is

$$\begin{aligned} \Pr(O_n \notin w | \Theta \notin \omega) &= 1 - \Pr(O_n \in w | \Theta \notin \omega) \\ &= 1 - P(w|\Theta) \text{ for } \Theta \notin \omega. \end{aligned}$$

The significance of the power of a test is now seen to be the following: For $\Theta \in \omega$, $P(w|\Theta)$ is the probability of committing a Type I error; for $\Theta \notin \omega$, $P(w|\Theta)$ is the probability of avoiding a Type II error. We illustrate this discussion with an example of a one parameter case*.

Suppose O_n is from $N(a,1)$, and that we wish to test the hypothesis

$$H_o: \quad a = a_o.$$

Let u_1, u_2 be any two numbers, $-\infty \leq u_1 < u_2 \leq +\infty$, such that

(b) $$(2\pi)^{-\frac{1}{2}} \int_{u_1}^{u_2} e^{-\frac{1}{2}u^2} du = 1 - \alpha.$$

Consider the test which consists of rejecting H_o if

(c) $$\sqrt{n}(\bar{x}-a_o) < u_1 \text{ or } \sqrt{n}(\bar{x}-a_o) > u_2.$$

The critical region w of the test is the part of the sample space defined by (c), that is, the region outside a certain pair of parallel hyperplanes (if $u_1 = -\infty$, or $u_2 = +\infty$, w is a half-space). Let us calculate the power of the test:

$$\begin{aligned} P(w|\Theta) &= \Pr[\sqrt{n}(\bar{x}-a_o) < u_1 \text{ or } \sqrt{n}(\bar{x}-a_o) > u_2 | a] \\ &= 1 - \Pr[u_1 \leq \sqrt{n}(\bar{x}-a_o) \leq u_2 | a]. \end{aligned}$$

Now if the true parameter value is $\underline{a}$,

$$u = \sqrt{n}(\bar{x}-a)$$

---

*An elementary discussion of a simple case with several parameters may be found in a paper by H. Scheffé, "On the ratio of the variances of two normal populations", Annals of Mathematical Statistics, Vol. 13 (1942), No. 4.

has the distribution N(0,1). Write

$$\sqrt{n}(\bar{x}-a_0) = u+\sqrt{n}(a-a_0).$$

Then

$$P(w|a) = 1-Pr[u_1-\sqrt{n}(a-a_0) \leq u \leq u_2-\sqrt{n}(a-a_0)|a],$$

$$\text{(d)} \qquad P(w|a) = 1-(2\pi)^{-\frac{1}{2}} \int_{u_1-\sqrt{n}(a-a_0)}^{u_2-\sqrt{n}(a-a_0)} e^{-\frac{1}{2}u^2}\, du.$$

Each choice of the pair of limits u_1, u_2 satisfying (b) gives a test of H_0. Let us now consider the class C of tests thus determined, and try to find which is the "best" test of the class C.

We note first that for all tests of the class,

$$Pr(\text{Type I error}) = P(w|a_0) = \alpha,$$

from (d) and (b). This is what we have previously called the significance level of the test. To compare the tests we might consider the graphs of $P(w|a)$ against $\underline{a}$ for the various tests; the graph for a given test is called the power curve of the test. We have seen that for every test of the class C, the power curve passes through the point (a_0,α). To find the shape of the power curve, we might plot points from (d), but by elementary methods we reach the following conclusions: The slope of the power curve corresponding to (u_1, u_2) is zero if and only if $\underline{a}$ is equal to

$$\text{(e)} \qquad a_m = a_0 + \tfrac{1}{2}(u_1+u_2)/\sqrt{n}.$$

As $a \longrightarrow +\infty$, $P(w|a) \longrightarrow 1$, unless $u_2 = +\infty$, in which case $P \longrightarrow 0$. As $a \longrightarrow -\infty$, again $P \longrightarrow 1$, unless $u_1 = -\infty$, in which case $P \longrightarrow 0$. Also, $0 < P < 1$. Except for the cases $u_1 = -\infty$ or $u_2 = +\infty$, the power curve must then behave as follows: rising from a minimum at the value $a = a_m$ given by (e), it increases monotonically, approaching the asymptote $P = 1$ as $a \longrightarrow \infty$. It may be shown from (d) and (e) that its behavior is symmetrical with respect to the line $a = a_m$. In the exceptional cases, for $u_1 = -\infty$, P increases monotonically from 0 to 1 as $\underline{a}$ increases from $-\infty$ to $+\infty$; for $u_2 = +\infty$, P decreases monotonically from 1 to 0. Some power curves are sketched in the figure:

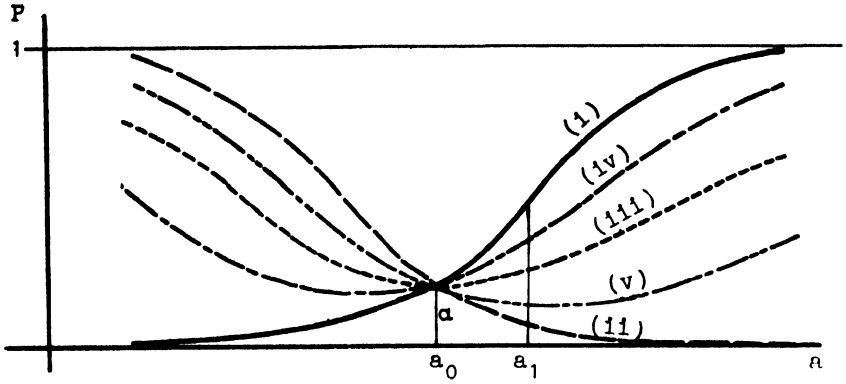

Figure 7

(i) is for the test with $u_1 = -\infty$, (ii) for $u_2 = +\infty$, (iii) has its minimum at a_0, (iv) to the left of a_0, (v) to the right. All the tests of class C have power curves lying in the region between the curves (i) and (ii).

As far as the probability of avoiding errors of Type I is concerned, the tests are equivalent, since the curves all pass through (a_0,α). For $a \neq a_0$, we recall that the ordinate on the curve is the probability of avoiding a Type II error. For two tests of H_0, say T_1 and T_2, with critical regions w_1 and w_2, we say that T_1 is more powerful than T_2 for testing H_0: $a = a_0$ against an alternative $a = a' \neq a_0$ if $P(w_1|a') > P(w_2|a')$. This means that if the true parameter value is a', the probability of avoiding a Type II error is greater in using T_1 than T_2. Now for altervatives $a > a_0$, the power curve (i) lies above all other power curves of tests of class C, that is, the test obtained by taking $u_1 = -\infty$ is the most powerful of the class C for all alternatives $a > a_0$. Hence this would be the best test of the class to use in a situation where we do not mind accepting H_0 if the true $a < a_0$, but want the most sensitive test of the class for rejecting H_0 when the true $a > a_0$. On the other hand, we see that this test is the worst of the lot, that is, the least powerful, for testing H_0 against alternatives $a < a_0$. For these alternatives the test with power curve (ii), obtained by taking $u_2 = +\infty$, is the most powerful. There is thus no test which is uniformly most powerful of the class C for all alternatives $-\infty < a < +\infty$.

The situation described in the last sentence is the common one. To deal with it Neyman and Pearson defined an unbiased test as one for which $P(w|a)$ is minimum for $a = a_0$. The argument against biased tests in a situation where we are interested in testing a hypothesis against all possible alternatives is that for a biased test Pr(accepting H_0) is greater if a has certain values $\neq a_0$ than if $a = a_0$. If we set $a_m = a_0$ in (e) we find that the unbiased test of the class C is that for which $u_1 = -u_2$. This is the test of the class C we should prefer, barring the "one-sided" situations where the tests with power curves (i) and (ii) are appropriate.

This serves to illustrate the comparison of tests by use of their power functions. Beyond this description of the underlying ideas of the Neyman-Pearson theory, it is not feasible to go into it further except for a few remarks: If one considers instead of the class C, the more inclusive class of all tests with critical regions w for which $P(w|a_0) = \alpha$, there is again no uniformly most powerful test. However, the unbiased test obtained above is actually the uniformly most powerful unbiased test of this broader class.

Leaving the one parameter case now, we recall that the definition of the power of a test and its meaning in terms of the probability of committing Type I and Type II errors was given for the multi-parameter case at the beginning of this section. Methods of finding optimum critical regions in the light of these concepts have been given by Neyman, Pearson, Wald and others, but there is still much work to be done. The problems of defining and finding "best" confidence intervals are related to those of "best" tests; the groundwork for such a theory has been laid by Neyman*. In conclusion, we recall the assumption made at the beginning of Chapter VII: that the functional form of the distribution is known for every possible parameter point: It is clear that in the application of the theory the calculations for the gain in efficiency by using a "best" test in preference to some other test will be invalidated if those calculations have been made for a distribution other than the true distribution. The whole theory introduced above presumes the knowledge of the functional form of the distribution.

*J. Neyman, "Outline of a theory of statistical estimation based on the classical theory of probability", Phil. Trans. Roy. Soc. London, Series A, Vol. 236 (1937), pp. 333-380.

CHAPTER VIII

NORMAL REGRESSION THEORY

In §2.9 certain ideas and definitions in regression theory were set forth and discussed. In the present chapter we shall consider sampling problems and tests of statistical hypotheses which arise in an important special type of regression theory which we shall refer to as normal regression theory. To be more specific, we shall assume that y is a random variable distributed according to $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$, where $x_1,\ldots,x_k$ are fixed variates, and consider samples of size n from such a distribution. $N(\sum_{1}^{k} a_p x_p, \sigma^2)$ is a conditional probability law of the form $f(y|x_1,x_2,\ldots,x_k)$. A sample of size n will consist of n sets of values $(y_\alpha|x_{1\alpha}, x_{2\alpha},\ldots,x_{k\alpha})$, $\alpha = 1,2,\ldots,n$, where $y_1,\ldots,y_n$ are n random variables, but where the $x_{p\alpha}$, $p = 1,2,\ldots,k$, $\alpha = 1,\ldots,n$, are fixed variates, and not random variables. We shall consider such problems as estimating (by confidence intervals and point estimation according to principles set forth in §6.1 and §6.2) values of the a's and σ^2 from the sample and of testing certain statistical hypotheses regarding the a's. We shall also consider applications of normal regression theory to certain problems in analysis of variance, including row-column and Latin square lay-outs.

8.1 Case of One Fixed Variate

In order to fix our ideas in the regression problem, we shall first consider in detail the case in which y is distributed according to $N(a+bx,\sigma^2)$. Let O_n: $(y_\alpha|x_\alpha)$, $\alpha = 1,2,\ldots,n$ $>k$, be a sample of size n from a population having this distribution. The probability element for the sample is

$$(a) \qquad dF(y_1,\ldots,y_n) = \prod_{\alpha=1}^{n} N(a+bx_\alpha,\sigma^2)dy_\alpha$$

$$= \left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}(y_\alpha-a-bx_\alpha)^2}\right]dy_1\ldots dy_n .$$

Maximizing the likelihood function (that enclosed in [ ]) with respect to σ^2, a, b, we find in accordance with §6.24 that $\hat{a}$ and $\hat{b}$ are given by solving

$$\text{(b)} \qquad \begin{aligned} \sum_1^n y_\alpha - \hat{a}n - \hat{b}\sum_1^n x_\alpha &= 0, \\ \sum_1^n x_\alpha y_\alpha - \hat{a}\sum_1^n x_\alpha - \hat{b}\sum_1^n x_\alpha^2 &= 0, \end{aligned}$$

and $\hat{\sigma}^2$ is given by

$$\text{(c)} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_1^n (y_\alpha - \hat{a} - \hat{b}x_\alpha)^2.$$

Solving (b) we obtain

$$\text{(d)} \qquad \begin{aligned} \hat{a} &= \bar{y} - \hat{b}\bar{x}, \\ \hat{b} &= \sum_1^n (x_\alpha - \bar{x})(y_\alpha - \bar{y}) \Big/ \sum_1^n (x_\alpha - \bar{x})^2. \end{aligned}$$

In order to be able to solve (b), we must have $\sum_1^n (x_\alpha - \bar{x})^2 \neq 0$. Now $\hat{a}$ and $\hat{b}$ are linear functions of $y_1, \ldots, y_n$, and it follows from Theorem (C) of §3.23 that $\hat{a}$ and $\hat{b}$ are jointly distributed according to a normal bivariate law with

$$E(\hat{b}) = b, \qquad E(\hat{a}) = a,$$

$$\sigma_{\hat{b}}^2 = \frac{\sigma^2}{\sum_1^n (x_\alpha - \bar{x})^2}, \qquad \sigma_{\hat{a}}^2 = \frac{\sigma^2 \sum_1^n x_\alpha^2}{n\sum_1^n (x_\alpha - \bar{x})^2},$$

$$\operatorname{cov}(\hat{a}, \hat{b}) = -\frac{\bar{x}\sigma^2}{\sum_1^n (x_\alpha - \bar{x})^2}.$$

The sum of squares in the exponent of (a) may be written as

$$\frac{1}{\sigma^2}\sum_1^n (y_\alpha - \hat{a} - \hat{b}x_\alpha + (\hat{a} - a) + (\hat{b} - b)x_\alpha)^2 = q_1 + q_2,$$

where

$$q_1 = \frac{1}{\sigma^2}\sum_1^n (y_\alpha - \hat{a} - \hat{b}x_\alpha)^2,$$

$$q_2 = \frac{1}{\sigma^2}\left[n(\hat{a} - a)^2 + 2\sum_1^n x_\alpha(\hat{a} - a)(\hat{b} - b) + \sum_1^n x_\alpha^2(\hat{b} - b)^2\right].$$

It is evident from (b) that $\hat{a} - a$, $\hat{b} - b$ are homogeneous linear functions of $(y_\alpha - a - bx_\alpha)$ $(\alpha = 1,2,\ldots,n)$. Also $y_\alpha - \hat{a} - \hat{b}x_\alpha = y_\alpha - a - bx_\alpha - (\hat{a}-a) - (\hat{b}-b)x_\alpha$ which is a homogeneous linear function of the $(y_\alpha - a - bx_\alpha)$ $(\alpha = 1,2,\ldots,n)$. We know that $\sum_1^n (y_\alpha - a - bx_\alpha)^2$ is distributed according to the χ^2-law with n degrees of freedom, and that q_2 (which is the exponent in the joint normal distribution of $\hat{a} - a$ and $\hat{b} - b$) is distributed according to a χ^2-law with 2 degrees of freedom. Therefore, it follows from Cochran's theorem §5.24 that q_1 is distributed according to the χ^2-law with $n - 2$ degrees of freedom.

We may summarize in the following

Theorem (A): Let O_n: $(y_\alpha | x_\alpha)$, $\alpha = 1,2,\ldots,n$, where the x_α are not all equal, be a sample of size n from a population with the distribution $N(a+bx,\sigma^2)$. Then

(1) The maximum likelihood estimates $\hat{a}$, $\hat{b}$ and $\hat{\sigma}^2$ of a, b, σ^2, respectively, are given by (b) and (c).

(2) $\hat{a} - a$ and $\hat{b} - b$ are jointly normally distributed with zero means and variance-covariance matrix given by

$$\left\| \begin{matrix} \frac{n}{\sigma^2} & \frac{\sum_1^n x_\alpha}{\sigma^2} \\ \frac{\sum_1^n x_\alpha}{\sigma^2} & \frac{\sum_1^n x_\alpha^2}{\sigma^2} \end{matrix} \right\|^{-1}$$

(3) $\frac{1}{\sigma^2}\sum_1^n (y_\alpha - \hat{a} - \hat{b}x_\alpha)^2$ is distributed according to the χ^2-law with $n - 2$ degrees of freedom and is distributed independently of $\hat{a}$ and $\hat{b}$.

One may readily set up confidence limits for a or b on basis of the "Student" t-distribution. For example, from §5.3 it follows that

$$t = \frac{\sqrt{\frac{(\hat{b}-b)^2}{\sigma^2}\sum_1^n (x_\alpha - \bar{x})^2}}{\sqrt{\frac{\frac{1}{\sigma^2}\sum_1^n (y_\alpha - \hat{a} - \hat{b}x_\alpha)^2}{n-2}}} = \frac{(\hat{b}-b)\sqrt{\sum_1^n (x_\alpha - \bar{x})^2}\sqrt{n-2}}{\sqrt{\sum_1^n (y_\alpha - \hat{a} - \hat{b}x_\alpha)^2}},$$

is distributed according to $g_{n-2}(t)$, from which confidence limits can be set up for b, or the statistical hypothesis can be tested that b has some specified value b_0 (e. g., 0, which corresponds to the hypothesis that y is independent of x). A similar treatment holds for a.

8.2 The Case of k Fixed Variates

Suppose y is distributed according to $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$. Let O_n: $(y_\alpha | x_{1\alpha}, x_{2\alpha}, \ldots, x_{k\alpha})$, $\alpha = 1,2,\ldots,n > k$, be a sample of size n from this distribution. The probability element for the sample is

$$\text{(a)} \qquad dF(y_1,\ldots,y_n) = \left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}(y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2}\right] dy_1 \ldots dy_n.$$

There is no loss of generality in considering the mean of y_α as a homogeneous linear function of $x_{1\alpha}, x_{2\alpha},\ldots,x_{k\alpha}$, for by choosing one of the x's, say $x_k = 1$ for all α, we can reduce our results so as to cover the case in which the mean value of y is not homogeneous in the fixed variates, i.e., of the form $(a_1 + a_2 x_2 + \ldots + a_k x_k)$. The results for the homogeneous case are simpler than for the non-homogeneous case from the point of view of notation because of greater symmetry.

The maximum likelihood estimates of $a_1,\ldots,a_k$ and σ^2, found by maximizing the quantity in [ ] in (a), are given by the following equations

$$\text{(b)} \qquad \sum_{\alpha=1}^{n} x_{q\alpha}(y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha}) = 0, \qquad (q = 1,2,\ldots,k)$$

$$\text{(c)} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{\alpha=1}^{n}(y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha})^2.$$

Let $a_{pq} = \sum_{\alpha=1}^{n} x_{p\alpha} x_{q\alpha}$, $a_{op} = \sum_{\alpha=1}^{n} y_\alpha x_{p\alpha}$, $a_{oo} = \sum_{1}^{n} y_\alpha^2$ $(p,q = 1,2,\ldots,k)$. We may write the equations (b) as

$$\text{(d)} \qquad a_{oq} - \sum_{p=1}^{k} \hat{a}_p a_{pq} = 0, \qquad (q = 1,2,\ldots,k).$$

If the determinant $|a_{pq}| \neq 0$, then it follows from §2.94 that the solution of (d) is

$$\text{(e)} \qquad \hat{a}_p = \sum_{q=1}^{k} a^{pq} a_{oq}.$$

It should be noted that $||a_{pq}||$ is positive definite and hence $|a_{pq}| \neq 0$ if $|x_{p\alpha}|$ are linearly independent (i. e., if there exists no set of real numbers C_p $(p = 1,2,\ldots,k)$ not all zero for which $\sum_{p=1}^{k} x_{p\alpha} C_p = 0$, for all α). For consider the quadratic form $\sum_{p,q=1}^{k} a_{pq} C_p C_q = \sum_{p,q=1}^{k} \sum_{\alpha=1}^{n} x_{p\alpha} x_{q\alpha} C_p C_q = \sum_{\alpha=1}^{n} (\sum_{p=1}^{k} x_{p\alpha} C_p)^2$. If the $x_{p\alpha}$ are linearly independent,

clearly $\sum_{\alpha=1}^{n}(\sum_{p=1}^{k} x_{p\alpha}c_p)^2$, and hence $\sum_{p,q=1}^{k} a_{pq}c_pc_q$, cannot vanish. Now the a_{op} and hence the $\hat{a}_p$ are linear functions of the random variables $y_1, y_2, \ldots, y_n$. Therefore, the $\hat{a}_p$ are distributed according to a normal k-variate distribution. The variance of $\hat{a}_p$ is $\sum_{\alpha=1}^{n}(\sum_{q} a^{pq}x_{q\alpha})^2 = \sum_{q,q'=1}^{k} a^{pq}a^{pq'}a_{qq'} = a^{pp}$. Similarly, the covariance of $\hat{a}_p$ and $\hat{a}_q$ is a^{pq}.

It will be noted that (b) can be written as

$$\sum_{\alpha=1}^{n} x_{q\alpha}[y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha} - \sum_{p=1}^{k}(\hat{a}_p - a_p)x_{p\alpha}] = 0, \qquad (q = 1,2,\ldots,k)$$

which shows that $(\hat{a}_p - a_p)$, $(p = 1,2,\ldots,k)$, are homogeneous linear functions of $y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha}$, $(\alpha = 1,2,\ldots,n)$. $y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha}$, $(\alpha = 1,2,\ldots,n)$ are also homogeneous linear functions of $y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha}$. Now

$$\frac{1}{\sigma^2}\sum_{\alpha=1}^{n}(y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2 = \frac{1}{\sigma^2}\sum_{\alpha=1}^{n}[(y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha}) + \sum_{p=1}^{k}(\hat{a}_p - a_p)x_{p\alpha}]^2 = q_1 + q_2,$$

where

(f)
$$q_1 = \frac{1}{\sigma^2}\sum_{\alpha=1}^{n}(y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha})^2,$$

$$q_2 = \frac{1}{\sigma^2}\sum_{p,q=1}^{k} a_{pq}(\hat{a}_p - a_p)(\hat{a}_q - a_q).$$

Hence, q_1 and q_2 are homogeneous quadratic forms in $(y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})$. Since $\hat{a}_p - a_p$ are distributed according to a k-variate normal law with variance-covariance matrix $\| \frac{a_{pq}}{\sigma^2} \|^{-1}$ it follows from §5.22 that q_2 is distributed according to the χ^2-law with k degrees of freedom. Similarly, we know that

$$\frac{1}{\sigma^2}\sum_{\alpha=1}^{n}(y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2 = q_1 + q_2 ,$$

is distributed according to the χ^2-law with n degrees of freedom. Therefore, by Cochran's Theorem, §5.24, q_1 is distributed according to the χ^2-law with n - k degrees of freedom and independently of q_2 (i. e., the $\hat{a}_p - a_p$).

Consider the sum of squares in (c); we may write

$$\frac{1}{\sigma^2}\sum_{\alpha=1}^{n}(y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha})^2 = \frac{1}{\sigma^2}(a_{oo} - 2\sum_{p=1}^{k} \hat{a}_p a_{op} + \sum_{p,q=1}^{k} \hat{a}_p \hat{a}_q a_{pq}).$$

But, it follows from §2.94 that this expression reduces to

$$(g) \qquad \frac{1}{\sigma^2 |a_{pq}|} \begin{vmatrix} a_{00} & a_{01} & \cdots & a_{0k} \\ a_{10} & a_{11} & \cdots & a_{1k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{k0} & a_{k1} & \cdots & a_{kk} \end{vmatrix} .$$

We may summarize in

Theorem (A): Let O_n: $(y_\alpha | x_{1\alpha}, x_{2\alpha}, \ldots, x_{k\alpha})$, $(\alpha = 1,2,\ldots,n)$, be a sample of size n from a population with distribution $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$, where $x_{1\alpha}, x_{2\alpha}, \ldots, x_{k\alpha}$, $(\alpha = 1,2,\ldots,n)$, are linearly independent. Let $a_{pq} = \sum_{\alpha=1}^{n} x_{p\alpha} x_{q\alpha}$, $a_{op} = \sum_{\alpha=1}^{n} y_\alpha x_{p\alpha}$, and $a_{oo} = \sum_{1}^{n} y_\alpha^2$. Then

(1) The maximum likelihood estimates of the a_p and σ^2 are given by (e) and (c).

(2) The quantities $(\hat{a}_p - a_p)$, $(p = 1,2,\ldots,k)$, are distributed according to a k-variable normal law with zero means and variance-covariance matrix $\| \frac{a_{pq}}{\sigma^2} \|^{-1}$.

(3) The quantity $\frac{1}{\sigma^2} \sum_{\alpha=1}^{n} (y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha})^2$ which may be expressed as in (g) as the ratio of two determinants, is distributed according to a χ^2-law with n - k degrees of freedom, and independently of the $(\hat{a}_p - a_p)$.

Making use of the results as stated in Theorem (A), one may set up confidence limits (or a significance test) for any a_p by setting up the appropriate Student ratio. Or one may set up confidence limits for σ^2 by using q_1. Confidence regions may be set up for all of the a_p or any sub-set of them by setting up a Snedecor F ratio, in which the numerator sum of squares is the exponent in the normal distribution of the corresponding set of $\hat{a}_p$ and the denominator sum of squares is q_1.

An Alternative Proof of the Independence of q_1 and q_2.

The proof which has been given for establishing the independence of the two above expressions in the probability sense depends upon Cochran's Theorem. The independence can also be established by the use of moment generating functions. Let $\phi(\theta_1, \theta_2)$ be the moment generating function defined as

$$\phi(\theta_1, \theta_2) = E(e^{\theta_1 q_1 + \theta_2 q_2}),$$

where q_1 and q_2 are defined in (f). If we can show that

$$\phi(\theta_1, \theta_2) = (1-2\theta_1)^{\frac{-(n-k)}{2}} (1-2\theta_2)^{-\frac{k}{2}},$$

then it follows by Theorem (B) in §2.81 and (e) of §3.3 that q_1 and q_2 are independently distributed according to χ^2-laws with n - k and k degrees of freedom respectively.

Let $\frac{1}{\sigma}(y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha}) = z_\alpha$. We may write equations (b) as

$$\sum_{\alpha=1}^{n} x_{q\alpha}(z_\alpha - \frac{1}{\sigma}\sum_{p=1}^{k}(\hat{a}_p - a_p)x_{p\alpha}) = 0, \qquad (q = 1,2,\ldots,k)$$

which reduces to

$$a'_{oq} - \frac{1}{\sigma}\sum_{p=1}^{k}(\hat{a}_p - a_p)a_{pq} = 0, \qquad (a'_{oq} = \sum_{\alpha=1}^{n} x_{q\alpha}z_\alpha)$$

from which we have

$$\frac{1}{\sigma}(\hat{a}_p - a_p) = \sum_{q=1}^{k} a^{pq}a'_{oq}.$$

Now

$$q_1 = \sum_{\alpha=1}^{n}(z_\alpha - \frac{1}{\sigma}\sum_{p=1}^{k}(\hat{a}_p - a_p)x_{p\alpha})^2 = \sum_{\alpha=1}^{n} z_\alpha^2 - \sum_{p,q=1}^{k} a'_{op}a'_{oq}a^{pq} = \sum_{\alpha=1}^{n} z_\alpha^2 - \sum_{\alpha,\beta=1}^{n} A_{\alpha\beta}z_\alpha z_\beta,$$

where

$$A_{\alpha\beta} = \sum_{p,q=1}^{k} a^{pq}x_{p\alpha}x_{q\beta}$$

For q_2 we have

$$q_2 = \frac{1}{\sigma^2}\sum_{p,q=1}^{k} a_{pq}(\hat{a}_p - a_p)(\hat{a}_q - a_q)$$
$$= \sum_{\alpha,\beta=1}^{n} A_{\alpha\beta}z_\alpha z_\beta.$$

The probability element associated with the sample is given by (a). Making the transformation $\frac{1}{\sigma}(y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha}) = z_\alpha$, $\alpha = 1,2,\ldots,n$, we obtain as the probability element of the z_α the expression

$$(\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum_{1}^{n} z_\alpha^2} dz_1 dz_2 \ldots dz_n.$$

For the m. g. f. we have

$$\phi(\theta_1,\theta_2) = (\frac{1}{\sqrt{2\pi}})^n \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-\frac{1}{2}Q} dz_1 dz_2 \ldots dz_n,$$

where

$$Q = \sum_{\alpha=1}^{n} z_\alpha^2 - 2\theta_1 q_1 - 2\theta_2 q_2,$$

$$= \sum_{\alpha=1}^{n} B_{\alpha\beta} z_\alpha z_\beta,$$

where $B_{\alpha\alpha} = 1 - 2\theta_1 + 2(\theta_1-\theta_2)A_{\alpha\alpha}$

and $B_{\alpha\beta} = 2(\theta_1-\theta_2)A_{\alpha\beta}, \qquad \alpha \neq \beta.$

The value of the n-tuple integral is $\frac{1}{\sqrt{B}}$ where B is the determinant $|B_{\alpha\beta}|$. To evaluate B, let us augment it as follows (letting $1 - 2\theta_1 = M$, $2(\theta_1-\theta_2) = N$):

$$B = \begin{vmatrix} M+NA_{11} & NA_{12} \cdots NA_{1n} & x_{11} & x_{21}\cdots x_{k1} \\ NA_{21} & M+NA_{22}\cdots NA_{2n} & x_{12} & x_{22}\cdots x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ NA_{n1} & NA_{n2} \cdots M+NA_{nn} & x_{1n} & x_{2n}\cdots x_{kn} \\ 0\cdots\cdots\cdots\cdots 0 & & 1 & 0\cdots\cdots 0 \\ \vdots & & 0 & 1\cdots\cdots 0 \\ 0\cdots\cdots\cdots\cdots 0 & & 0 & 0\cdots\cdots 1 \end{vmatrix}.$$

Suppose the (n+p)-th column is multiplied by $-N(\sum_{q=1}^{k} a^{pq}x_{q\alpha}) = C_p$, say, $p = 1,2,\ldots,k$, and added to the α-th column $\alpha = 1,2,\ldots,n$. We obtain the following expression for B

$$B = \begin{vmatrix} M & 0 \cdots\cdots 0 & x_{11} & x_{21}\cdots x_{k1} \\ 0 & M \cdots\cdots 0 & x_{12} & x_{22}\cdots x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0\cdots\cdots\cdots M & x_{1n} & x_{2n}\cdots x_{kn} \\ C_{11} & C_{12}\cdots\cdots C_{1n} & 1 & 0\cdots\cdots 0 \\ C_{21} & C_{22}\cdots\cdots C_{2n} & 0 & 1\cdots\cdots 0 \\ \vdots & \vdots & \vdots & \vdots \\ C_{k1} & C_{k2}\cdots\cdots C_{kn} & 0 & 0\cdots\cdots 1 \end{vmatrix}.$$

Now suppose the α-th column ($\alpha = 1,2,\ldots,n$) be multiplied by $-\frac{x_{p\alpha}}{M}$ and added to the (n+p)-th column ($p = 1,2,\ldots,k$). Noting that

$$-\sum_{\alpha=1}^{n} \frac{C_{q\alpha}x_{p\alpha}}{M} = \frac{N}{M}\sum_{\alpha=1}^{n}\sum_{q=1}^{k} a^{qq'}x_{q\alpha}x_{p\alpha} = \frac{N}{M}\sum_{q'=1}^{k} a^{qq'}a_{pq'} = \begin{cases} 0 & p\neq q \\ 1 & p=q \end{cases}.$$

We find that B reduces to

$$B = \begin{vmatrix} M & 0 \cdots 0 & 0 & 0 \cdots\cdots 0 \\ 0 & M \cdots 0 & 0 & 0 \cdots\cdots 0 \\ \vdots & \vdots\ddots\vdots & \vdots & \vdots\ddots\vdots \\ 0 & 0 \cdots M & 0 & 0 \cdots\cdots 0 \\ C_{11} & C_{12} \cdots C_{1n} & (1+\tfrac{N}{M}) & 0 \cdots\cdots 0 \\ C_{21} & C_{22} \cdots C_{2n} & 0 & (1-\tfrac{N}{M}) \cdots 0 \\ \vdots & \vdots\ddots\vdots & \vdots & \vdots\ddots\vdots \\ C_{k1} & C_{k2} \cdots C_{kn} & 0 & 0 \cdots\cdots (1-\tfrac{N}{M}) \end{vmatrix}$$

$$= M^{n-k}(M+N)^{k}.$$

Therefore we have

$$\phi(\theta_1,\theta_2) = \frac{1}{\sqrt{B}} = (1-2\theta_1)^{-\frac{n-k}{2}}(1-2\theta_2)^{-\frac{k}{2}},$$

which concludes the argument that q_1 and q_2 are independently distributed according to χ^2-laws with n - k and k degrees of freedom respectively.

Remarks on the Generality of the linear Regression Function. The regression function $\sum_{p=1}^{k} a_p x_{p\alpha}$ is much more general than it might appear at first. For example, if $x_1 = 1$, $x_2 = t$, $x_3 = t^2$, ..., $x_k = t^{k-1}$, the regression function would be the polynomial $\sum_{p=1}^{k} a_p t^{p-1}$, in which case we would have a random variable y, having as its mean value the function of t given by $\sum_{p=1}^{k} a_p t^{p-1}$. The estimates $\hat{a}_p$ would, of course, have the same form as those given by (e), except that $x_{p\alpha}$ would be replaced by t_α^{p-1} in calculating a_{pq} and a_{oq}.

Again, we might have for $k = 2m+1$, $x_1 = 1$, $x_2 = \sin t$, $x_3 = \cos t$, $x_4 = \sin 2t$, ..., $x_{2m+1} = \cos mt$, in which case the mean value of y is a harmonic function of the form $a_1 + a_2 \sin t + a_3 \cos t + \cdots + a_{2m+1} \cos mt$. The procedure for obtaining $\hat{a}_1$, $\hat{a}_2$,..., $\hat{a}_{2m+1}$ is as before given by (e).

Another example: Suppose $k = 2$ and $x_{1\alpha} = 1$, $x_{2\alpha} = 0$ for $\alpha = 1,2,\ldots,n_1$, and $x_{1\alpha} = 0$, $x_{2\alpha} = 1$ for $\alpha = n_1+1,\ldots,n_1+n_2$. The sample O_n: $(y_\alpha | x_{1\alpha}, x_{2\alpha})$, $\alpha = 1,2,\ldots,n=n_1+n_2$ drawn from $N(\sum_{p=1}^{2} a_p x_p, \sigma^2)$ is equivalent to two independent samples O_{n_1}: $(y_1, y_2, \ldots, y_{n_1})$, O_{n_2}: $(y_{n_1+1}, \ldots y_{n_1+n_2})$ of size n_1 and n_2 respectively, drawn from $N(a_1,\sigma^2)$ and $N(a_2,\sigma^2)$ respectively. This example extends readily to the case of several independent samples.

Curvilinear regression is also a special case. For example, for quadratic regression in two variables, say u and v, we would let $x_1 = 1$, $x_2 = u$, $x_3 = v$, $x_4 = u^2$, $x_5 = v^2$, $x_6 = uv$.

8.3 A General Normal Regression Significance Test

The following general significance test frequently arises in normal regression theory: A sample O_n: $(y_\alpha | x_{1\alpha}, x_{2\alpha}, \ldots, x_{k\alpha})$, $\alpha = 1,2,\ldots,n$, is assumed to be drawn from a population with distribution $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$, and it is desired to test the hypothesis that a_{r+1}, a_{r+2}, ..., a_k ($r<k$) have specified values, say $a_{r+1,0}$, $a_{r+2,0}$, ..., $a_{k,0}$, respectively, no matter what values a_1, a_2, ..., a_r and σ^2 may have. For example, all specified values may be zero in which case the problem is to test the hypothesis that y, which is assumed to be distributed according to $N(\sum_{1}^{k} a_p x_p, \sigma^2)$, is actually independent of x_{r+1}, x_{r+2}, ..., x_k.

In order to determine the test function (of the y_α and $x_{p\alpha}$) for testing the hypothesis we shall make use of the method of likelihood ratios discussed in §7.2.

The probability element of the sample is

$$\text{(a)} \qquad dF(y_1,\ldots y_n) = P(O_n; a_1, a_2, \ldots, a_k, \sigma^2)\, dy_1 \ldots dy_n,$$

where

$$\text{(b)} \qquad P(O_n, a_1, a_2, \ldots, a_k, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}(y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2}$$

is the likelihood function.

Let Ω be the $(k+1)$-dimensional parameter space for which $\sigma^2 > 0$, $-\infty < a_p < +\infty$, $p = 1,2,\ldots,k$, and let ω be the $(k-r)$-dimensional subspace of Ω for which $a_{r+1} = a_{r+1,0}$, $a_{r+2} = a_{r+2,0}$, ..., $a_k = a_{k,0}$. If Ho denotes the hypothesis to be tested, then Ho is the hypothesis that the true parameter point lies in ω, where the admissible points are those in Ω.

The likelihood ratio λ for testing Ho is given by

$$\text{(c)} \qquad \lambda = \frac{\max_\omega P(O_n; a_1, \ldots, a_k, \sigma^2)}{\max_\Omega P(O_n; a_1, \ldots, a_k, \sigma^2)},$$

where the denominator is the maximum of $P(O_n; a_1, \ldots, a_k, \sigma^2)$ for variations of the parameters over Ω and the numerator is the maximum for variations of the parameters over ω. To find the maximum of the likelihood function (b) over Ω, we follow the ordinary procedure of taking the first derivative of the likelihood function with respect to each parameter,

setting the derivatives equal to zero. We find that the maximizing values are

(d) $$\hat{a}_p = \sum_{q=1}^{k} a^{pq} a_{oq}, \qquad \hat{\sigma}_\Omega^2 = \frac{1}{n}\sum_{\alpha=1}^{n} (y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha})^2,$$

$p = 1,2,\ldots,k$, as given in §8.2. Substituting these in the likelihood function we find

$$\max_\Omega P(O_n; a_1, a_2, \ldots, a_k, \sigma^2) = (\frac{1}{2\pi\hat{\sigma}_\Omega^2})^{\frac{n}{2}} e^{-\frac{n}{2}}.$$

Similarly, by maximizing the likelihood over ω, we set $a_{r+1} = a_{r+1,0}, \ldots, a_k = a_{k,0}$, and differentiate with respect to $\sigma^2, a_1, \ldots, a_r$, obtaining

(e) $$\hat{a}_u = \sum_{v=1}^{r} \dot{a}^{uv} \dot{a}_{ov}, \qquad \hat{\sigma}_\omega^2 = \frac{1}{n}\sum_{\alpha=1}^{n} (\dot{y}_\alpha - \sum_{u=1}^{k} \hat{a}_u x_{u\alpha})^2,$$

where $\dot{y}_\alpha = y_\alpha - \sum_{g=r+1}^{k} a_{g,0} x_{g\alpha}$, $\dot{a}_{uv} = \sum_{\alpha=1}^{n} x_{u\alpha} x_{v\alpha}$, $\dot{a}_{ou} = \sum_{\alpha=1}^{n} \dot{y}_\alpha x_{u\alpha}$, $\dot{a}_{oo} = \sum_{\alpha=1}^{n} \dot{y}_\alpha^2$, and $||\dot{a}^{uv}|| = ||\dot{a}_{uv}||^{-1}$, $u,v = 1,2,\ldots,r$. Substituting in the likelihood function, we find

$$\max_\omega P(O_n; a_1, \ldots, a_k, \sigma^2) = (\frac{1}{2\pi\hat{\sigma}_\omega^2})^{\frac{n}{2}} e^{-\frac{n}{2}}.$$

Therefore

(f) $$\lambda = (\frac{\hat{\sigma}_\Omega^2}{\hat{\sigma}_\omega^2})^{\frac{n}{2}}.$$

Now it is clear that $\hat{\sigma}_\Omega^2 \leq \hat{\sigma}_\omega^2$, since $\hat{\sigma}_\Omega^2$ is the minimum of $\frac{1}{n}\sum_{\alpha=1}^{k} (y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2$ for variations of $a_1, \ldots, a_k$, while $\hat{\sigma}_\omega^2$ is the minimum for variations of $a_1, \ldots, a_r$, for fixed values of $a_{r+1}, \ldots, a_k$. Now let

$$q_1 = \frac{n\hat{\sigma}_\Omega^2}{\sigma^2} \text{ and } q_2 = \frac{n(\hat{\sigma}_\omega^2 - \hat{\sigma}_\Omega^2)}{\sigma^2}.$$

The difference $n(\hat{\sigma}_\omega^2 - \hat{\sigma}_\Omega^2)$ is simply the further reduction in the sum of squares $\sum_{\alpha=1}^{n} (y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2$ obtainable by varying $a_{r+1}, \ldots, a_k$ in addition to $a_1, a_2, \ldots, a_r$. Expressed in terms of q_1 and q_2,

(g) $$\lambda = (\frac{q_1}{q_2+q_1})^{\frac{n}{2}} = (\frac{1}{1+q_2/q_1})^{\frac{n}{2}}.$$

Thus λ is a single-valued function of q_2/q_1, which means that q_2/q_1 is equivalent to λ as a test function. The nearer the value of λ to unity, the smaller the value of q_2 as

compared with q_1. To complete the problem of setting up a test for testing H_0, we must now obtain the distribution of λ, (or q_2/q_1) under the assumption that the hypothesis H_0 is true, i. e., that O_n has been drawn from $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$ for $a_{r+1} = a_{r+1,0}, \ldots, a_k = a_{k,0}$.

We shall now show that if H_0 is true then q_1 and q_2 are distributed independently according to χ^2-laws with $n - k$ and $k - r$ degrees of freedom respectively.

The probability element for the sample O_n from the population having distribution $N(\sum_{1}^{k} a_p x_p, \sigma^2)$ is given by (a). Now as we have seen in §8.2, the sum of squares in the exponent can be written as

$$\text{(h)} \qquad \sum_{\alpha=1}^{n}(y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha})^2 + \sum_{p,q=1}^{k} a_{pq}(\hat{a}_p - a_p)(\hat{a}_q - a_q).$$

The second expression in (h) may be written as

$$\text{(i)} \qquad \sum_{u,v=1}^{r} a_{uv}(\hat{a}_u - a_u + L_u)(\hat{a}_v - a_v + L_v) + \sum_{g,h=r+1}^{k} b_{gh}(\hat{a}_g - a_g)(\hat{a}_h - a_h),$$

where L_u $(u = 1,2,\ldots,r)$ are linear functions of $(\hat{a}_g - a_g)$, $(g = r+1,\ldots,k)$ and where $||b_{gh}|| = ||a^{gh}||^{-1}$, $g,h = r+1,\ldots,k$, where a^{gh} is the element in the g-th row and h-th column in $||a_{pq}||^{-1}$ $p,q = 1,2,\ldots,k$. See §3.23.

To verify the statement that expression (i) is equal to the second expression in (h), let us denote $\hat{a}_u - a_u$ by d_u and let $L_u = \sum_{g=r+1}^{k} l_{ug} d_g$. We must then determine the l_{ug} and the b_{gh} so that

$$\text{(j)} \qquad \sum_{u,v=1}^{r} a_{uv}(d_u + \sum_{g=r+1}^{k} l_{ug} d_g)(d_v + \sum_{h=r+1}^{k} l_{vh} d_h) + \sum_{g,h=r+1}^{k} b_{gh} d_g d_h \equiv \sum_{p,q=1}^{k} a_{pq} d_p d_q,$$

that is, an identity in the d's.

Taking $\frac{\partial^2}{\partial d_g \partial d_v}$ of both sides of this identity we get

$$\text{(k)} \qquad \sum_{u=1}^{r} a_{uv} l_{ug} = a_{vg}, \qquad \begin{pmatrix} v = 1,2,\ldots,r \\ g = r+1,\ldots,k \end{pmatrix}$$

and hence

$$\text{(l)} \qquad l_{ug} = \sum_{v=1}^{r} a^{uv}_{(r)} a_{vg} \text{ where } ||a^{uv}_{(r)}|| = ||a_{uv}||^{-1}.$$

Taking $\frac{\partial^2}{\partial d_g \partial d_h}$ of both sides of (j) we get

(m) $$\sum_{u,v=1}^{r} a_{uv} l_{ug} l_{vh} + b_{gh} = a_{gh}.$$

Using (k) and (l) we find that (m) reduces to

(n) $$\sum_{u,v=1}^{r} a^{uv}_{(r)} a_{uh} a_{vg} + b_{gh} = a_{gh},$$

or

(o) $$b_{gh} = a_{gh} - \sum_{u,v=1}^{r} a^{uv}_{(r)} a_{uh} a_{vg}.$$

Referring to §2.94 it will be seen that

(p) $$b_{gh} = \frac{1}{|a_{uv}|} \cdot \begin{vmatrix} a_{11} & \cdots & a_{1r} & a_{1h} \\ \vdots & \ddots & \vdots & \vdots \\ a_{r1} & \cdots & a_{rr} & a_{rh} \\ a_{g1} & \cdots & a_{gr} & a_{gh} \end{vmatrix}$$

which is equal to the term in the g-th row and h-th column in the inverse of $||a_{pq}||$.

Making use of the fact that the sum of the squares in the exponent of the likelihood function in (a) is

(q) $$\sum_{\alpha=1}^{n} (y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha})^2 + \text{expression (i)},$$

it is now clear that by maximizing the likelihood function for variations of $a_1,\ldots,a_k$, σ^2 in ω, we find

(r) $$\hat{\sigma}^2_\omega = \frac{1}{n}[(\sum_{\alpha=1}^{n} (y_\alpha - \sum_{p=1}^{k} \hat{a}_p x_{p\alpha})^2 + \sum_{g,h=r+1}^{k} b_{gh}(\hat{a}_g - a_{g,0})(\hat{a}_h - a_{h,0})],$$

since the first expression in (i) vanishes when $a_1,\ldots,a_r$ are varied so as to maximize the likelihood function.

Remembering that when the likelihood function is maximized with respect to a_1, $a_2,\ldots,a_k,\sigma^2$ over Ω we obtain $\hat{\sigma}^2_\Omega$ as given in (d), we clearly have

(s) $$q_2 = \frac{n(\hat{\sigma}^2_\Omega - \hat{\sigma}^2_\omega)}{\sigma^2} = \frac{1}{\sigma^2} \sum_{g,h=r+1}^{k} b_{gh}(\hat{a}_g - a_{g,0})(\hat{a}_h - a_{h,0}),$$

which is a function of the $\hat{a}_g$ $(g = r+1,\ldots,k)$ which, as we have seen in §8.2, are distri-

buted independently of $\frac{n\hat{\sigma}_\Omega^2}{\sigma^2} = q_1$, q_1 being distributed according to the χ^2-law with n - k degrees of freedom. But $-\frac{1}{2}q_2$ is seen to be the exponent in the joint distribution of $\hat{a}_g$ when H_o is true. Hence by §5.23 q_2 is distributed according to the χ^2-law with k - r degrees of freedom.

Since q_2 and q_1 are distributed independently according to χ^2-laws with k - r and n - k degrees of freedom when H_o is true it follows that the quantity

$$F = \frac{q_2}{q_1}\frac{(n-k)}{(k-r)},$$

is a Snedecor ratio distributed according to the law $h_{k-r,n-k}(F)dF$ (see §5.4) when H_o is true. Now

$$\lambda = \left(\frac{1}{1+q_2/q_1}\right)^{\frac{n}{2}} = \left(\frac{1}{1+\frac{k-r}{n-k}F}\right)^{\frac{n}{2}},$$

which shows that λ is a single-valued function of F, and hence F is equivalent to λ for testing the hypothesis H_o, the upper tail of the F-distribution being used for determining critical values of F for various significance levels. It can be shown that this test is unbiased (see §7.3) although we shall not demonstrate this fact here.*

We may summarize our results in the following fundamental Theorem in normal regression theory:

Theorem (A): Let O_n: $(y_\alpha|x_{1\alpha},x_{2\alpha},\ldots,x_{k\alpha})$ be a sample from a population with distribution $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$, where x_p are linearly independent. Let H_o be the statistical hypothesis that the true parameter point $a_1, a_2,\ldots, a_k, \sigma^2$ belongs to ω: $-\infty < a_u < +\infty$, $(u = 1,2,\ldots,r)$, $\sigma^2 > 0$, $a_{r+1} = a_{r+1,0}\ldots, a_k = a_{k,0}$ which is a subset of the admissible set Ω: $-\infty < a_p < +\infty$, $p = 1,2,\ldots,k$ $\sigma^2 > 0$. Let $\hat{\sigma}_\Omega^2$ be the maximum likelihood estimate of σ^2 for variations of the parameters over Ω, and $\hat{\sigma}_\omega^2$ the maximum likelihood estimate of σ^2 for variations of the parameters over ω. Then:

(1) $n\hat{\sigma}_\Omega^2 = \sum_{\alpha=1}^{n}(y_\alpha - \sum_{p=1}^{k}\hat{a}_p x_{p\alpha})^2$,

(2) $n\hat{\sigma}_\omega^2 = n\hat{\sigma}_\Omega^2 + \sum_{g,h=r+1}^{k} b_{gh}(\hat{a}_g - a_{g,o})(\hat{a}_h - a_{h,o})$ where $\hat{a}_p$ are given by (d), $||b_{gh}||$ is the inverse of the matrix obtained by deleting the first r rows and columns of $||a_{pq}||^{-1}$ $p,q = 1,2,\ldots,k$.

(3) The quantities

$$q_1 = \frac{n\hat{\sigma}_\Omega^2}{\sigma^2}, \qquad q_2 = \frac{n(\hat{\sigma}_\omega^2 - \hat{\sigma}_\Omega^2)}{\sigma^2},$$

are independently distributed according to χ^2-laws with n - k and k - r degrees of

*See J. F. Daly, "On the Unbiased Character of Likelihood Ratio Tests for Independence in Normal Systems", Annals of Math. Stat., Vol. 11, (1940).

freedom, respectively, when H_o is true.

(4) The likelihood criterion λ for testing H_o is given by

$$\lambda = \left(\frac{1}{1+q_2/q_1}\right)^{\frac{n}{2}} = \left(\frac{1}{1+\frac{k-r}{n-k}F}\right)^{\frac{n}{2}},$$

where F is Snedecor's ratio which is distributed according to

$$h_{k-r,n-k}(F)dF,$$

when H_o is true.

8.4 Remarks on the Generality of Theorem (A), §8.3

In order to emphasize the generality of Theorem (A), §8.3, we shall discuss briefly several cases of particular interest.

8.41 Case 1. It frequently happens that the following statistical hypothesis is to be tested on basis of a sample O_n: $(y_\alpha | x_{1\alpha}, x_{2\alpha}, \ldots, x_{k\alpha})$ assumed to have been drawn from a population with distribution $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$:

$$\Omega: \quad \{ -\infty < a_p < +\infty, \qquad \sigma^2 > 0, \qquad p = 1,2,\ldots,k$$

$$\omega: \quad \{ \text{Region in } \Omega \text{ for which } \sum_{p=1}^{k} c_{up} a_p = 0, \qquad \sigma^2 > 0, \quad u = r+1, r+2, \ldots, k$$

where the c_{up} are linearly independent constants. In other words the hypothesis to be tested here is that there are k - r linear restrictions among the a_p, given that the sample is from a population with distribution $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$. Denoting this statistical hypothesis by H_o', we may verify from Theorem (A) that the likelihood criterion for testing H_o' is of the same form as λ for H_o where $\sigma^2 q_1$ is the minimum of $S = \sum_{\alpha=1}^{n} (y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2$ for variations of $a_1, \ldots, a_p$ over Ω, and $\sigma^2 q_2$ is the difference between the minimum of S over ω and the minimum of S over Ω. As in the case of H_o, q_1 and q_2 for H_o' are independently distributed according to χ^2-laws with n - k and k - r degrees of freedom respectively, when H_o' is true. To verify that this is true, we transform the a_p as follows:

$$a_u = a_u', \qquad u = 1,2,\ldots, r$$

$$\sum_{p=1}^{k} c_{gp} a_p = a_g', \qquad g = r+1, \ldots, k.$$

We may write this transformation as

$$\sum_{p=1}^{k} c_{qp} a_p = a_q', \qquad q = 1,2,\ldots,k,$$

where

$$c_{gp} = 1 \qquad g = p \leq r$$
$$= 0 \qquad \text{otherwise} .$$

Without loss of generality we may assume the c_{gp} to be such that $|c_{qp}| \neq 0$. Hence

$$a_p = \sum_{q=1}^{k} c^{qp} a'_q ,$$

and $\sum_{p=1}^{k} a_p x_p = \sum_{q=1}^{k} a'_q x'_q$ where $x'_q = \sum_{p=1}^{k} c^{qp} x_p$. Therefore H'_o may be expressed as the statistical hypothesis:

$$\Omega: \ -\infty < a'_p < \infty, \qquad \sigma^2 > 0 \qquad p = 1,2,\ldots,k,$$
$$\omega: \text{ Region in } \Omega \text{ for which } a'_g = 0, \qquad \sigma^2 > 0 \qquad g = r+1,\ldots,k,$$

which is to be tested on basis of the sample O_n: $(y_\alpha | x'_{1\alpha}, x'_{2\alpha}, \ldots, x'_{k\alpha})$ $\alpha = 1,2,\ldots,n$ drawn from a population with distribution $N(\sum_{1}^{k} a'_p x'_p, \sigma^2)$. Theorem (A) is immediately applicable to H'_o as expressed in this form.

8.42 Case 2. The following statistical hypothesis say H''_o frequently arises, to be tested on basis of a sample O_n from $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$:

$$\Omega: \ -\infty < a_p < \infty, \qquad \sigma^2 > 0, \ p = 1,2,\ldots,k$$
$$\omega: \text{ Region in } \Omega \text{ for which } a_p = \sum_{u=1}^{r} c^{up} a'_u, \quad \sigma^2 > 0, \ p = 1,2,\ldots,k.$$

In other words the hypothesis H''_o is that the a_p can each be expressed linearly in terms of r ($<k$) parameters $a'_1,\ldots,a'_r$ where the c^{up} are given. By using the transformation $a_p = \sum_{q=1}^{k} c^{qp} a'_q$, where the c^{qp}, $(q = r+1,\ldots,k,\ p = 1,2,\ldots,k)$ are further given numbers such that $|c^{qp}| \neq 0$ we can express H''_o as follows:

$$\Omega: \ -\infty < a'_p < \infty, \qquad \sigma^2 > 0, \qquad p = 1,2,\ldots,k$$
$$\omega: \qquad a'_g = 0, \qquad \sigma^2 > 0, \qquad g = r+1,\ldots,k,$$

to be tested on basis of the sample O_n: $(y_\alpha | x'_{1\alpha},\ldots,x'_{k\alpha})$ $\alpha = 1,2,\ldots,n$, $x'_q = \sum_{p=1}^{k} c^{qp} x_p$ from a population with distribution $N(\sum_{p=1}^{k} a'_p x'_p, \sigma^2)$. This case is clearly covered by Theorem (A).

In this case $\sigma^2 q_1$ is the minimum of $S = \sum_{\alpha=1}^{n} (y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2$ for variations of a_1, $a_2,\ldots,a_k$ over Ω, while $\sigma^2 q_2$ is the difference between the minimum of S for variations of the a_p over Ω and that of S for variations of the a_p over ω(i.e., for unrestricted variations of a'_u, $u=1,2,\ldots,r$, when the a_p are replaced by $\sum_{q=1}^{k} c^{qp} a'_q$ in S and the a'_g are set equal to 0 $(g=r+1,\ldots,k)$.

q_1 and q_2 are independently distributed according to χ^2-laws with n - k and k - r degrees of freedom, respectively, when H_0'' is true.

8.43 Case 3. The following variant of the hypothesis H_0, of §8.3 arises in such problems as randomized blocks (see § 9.2), Latin squares (see § 9.4), etc., to be tested on basis of a sample 0_n: $(y_\alpha | x_{1\alpha},\ldots,x_{k\alpha})$ $\alpha = 1,2,\ldots,n$. Denoting this hypothesis by H_0''', it is specified as follows:

$$\Omega: \begin{cases} -\infty < a_p < +\infty, \qquad \sigma^2 > 0, \qquad p = 1,2,\ldots,k, \\ \text{with the } a_p \text{ restricted by the } r_1 \text{ linearly independent conditions.} \\ \sum_{p=1}^{k} d_{p\mu} a_p = 0, \qquad \mu = 1,2,\ldots,r_1 < k, \end{cases}$$

$$\omega: \begin{cases} \text{The subspace in } \Omega \text{ for which} \\ \sum_{p=1}^{k} d_{p\nu} a_p = 0, \qquad \nu = 1,2,\ldots,r_2, \text{ where } r_1 < r_2 < k. \end{cases}$$

H_0''' is the hypothesis that the a_p satisfy $r_2 - r_1$ further linear restrictions, assuming that r_1 linear restrictions are fulfilled, linear independence being assumed throughout. H_0''' is to be tested on basis of a sample 0_n: $(y_\alpha | x_{1\alpha},\ldots, x_{k\alpha})$ $\alpha = 1,2,\ldots,n$. In this case $\sigma^2 q_1$ is the minimum of $S = \sum_{\alpha=1}^{n} (y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2$ for variations of the a_p over Ω (i. e. for variations of the a_p subject to the restrictions $\sum_{p=1}^{k} d_{p\mu} a_p$, $\mu = 1,2,\ldots,r_1$) while $\sigma^2 q_2$ is the difference between this minimum and that for variations of the a_p over ω (i. e. for variations of the a_p subject to the restriction $\sum_{p=1}^{k} d_{p\nu} a_p$, $\nu = 1,2,\ldots,r_2$). When H_0''' is true, q_1 and q_2 are independently distributed according to χ^2-laws with $n - k - r_1$ and $r_2 - r_1$ degrees of freedom respectively.

That this case is covered by Theorem (A) may be seen by considering the non-singular transformation of the a_p, $\sum_{p=1}^{k} d_{pq} a_p = a'_q$, $q = 1,2,\ldots,k$ where d_{pq} are given numbers since that $|d_{pq}| \neq 0$. We have $a_p = \sum_{q=1}^{k} d^{pq} a'_q$ which transforms $\sum_{p=1}^{k} a_p x_{p\alpha}$ into $\sum_{p=1}^{k} a'_p x'_{p\alpha}$, where $x'_{p\alpha} = \sum_{p=1}^{k} d^{pq} x_{p\alpha}$. Now under Ω the regression function is $\sum_{p=r_1+1}^{k} a'_p x'_{p\alpha}$, and we may therefore specify H_0''' as

$$\Omega: \begin{cases} -\infty < a'_p < \infty, \ \sigma^2 > 0, \ p = r_1+1,\ldots,k \ (a'_1,\ldots,a'_{r_1} \\ \text{being assumed 0 from the outset)} \end{cases}$$

$$\omega: \begin{cases} \text{Subspace in } \Omega \text{ for which} \\ a'_{r_1+1} = \ldots = a'_{r_2} = 0. \end{cases}$$

The applicability of Theorem (A) is now obvious.

8.5 The Minimum of a Sum of Squares of Deviations with Respect to Regression Coefficients which are Subject to Linear Restrictions

It will be noted in §§8.3 and 8.4 that frequently we have to find the minimum of

(a) $$S = \sum_{\alpha=1}^{n} (y_\alpha - \sum_{p=1}^{k} a_p x_{p\alpha})^2,$$

for variations of the a_p, when the a_p are subject to one or more linear restrictions. The object of this section is to give an explicit expression for the minimum of the sum of squares, under such conditions, as a ratio of two determinants.

Let us consider the problem of finding the minimum of the sum of squares (a), when the a_p are subject to the linear restrictions

(b) $$\sum_{p=1}^{k} c_{up} a_p = 0, \qquad (u = 1,2,\ldots,r < k).$$

We shall use the method of Lagrange, §4.7 , and write

(c) $$F(a_1,a_2,\ldots,a_k;\lambda_1,\lambda_2,\ldots,\lambda_r) = S + 2\sum_{u=1}^{r} \lambda_u (\sum_{p=1}^{k} c_{up} a_p).$$

It is necessary that

(d) $$0 = \frac{\partial F}{\partial a_q} = \frac{\partial S}{\partial a_q} + 2\sum_{u=1}^{r} \lambda_u c_{uq} = 0, \qquad (q = 1,2,\ldots,k)$$

in order for S to have an extremum (in this case a minimum). Performing the differentiation $\frac{\partial S}{\partial a_q}$, these equations may be written as

(e) $$-a_{oq} + \sum_{p=1}^{k} a_p a_{pq} + \sum_{u=1}^{r} \lambda_u c_{uq} = 0, \qquad (q = 1,2,..,k),$$

where a_{oq}, a_{pq} (and a_{oo}) are defined in (d) of §8.2. Multiplying each of (e) by a_q and summing from q = 1 to k, we get

(f) $$-\sum_{q=1}^{k} a_{oq} a_q + \sum_{p,q=1}^{k} a_{pq} a_p a_q = 0.$$

Expanding S, we obtain

(g) $$S = a_{oo} - 2\sum_{p=1}^{k} a_{op} a_p + \sum_{p,q=1}^{k} a_{pq} a_p a_q,$$

and making use of (f),

(h) $$\sum_{p=1}^{k} a_{op} a_p = a_{oo} - S.$$

Rewriting (h) and (e) with $a_0 = 1$, and using the conditions (b), we obtain the following homogeneous linear equations in the $1 + k + r$ quantities $a_0, a_1, a_2, \ldots, a_k, \lambda_1, \ldots, \lambda_r$.

$$(S-a_{00})a_0 + \sum_{p=1}^{k} a_{0p}a_p = 0,$$

$$\text{(i)} \qquad -a_{q0}a_0 + \sum_{p=1}^{k} a_{pq}a_p + \sum_{u=1}^{r} c_{uq}\lambda_u = 0, \qquad q = 1,2,\ldots,k$$

$$\sum_{p=1}^{k} c_{up}a_p = 0, \qquad u = 1,2,\ldots,r.$$

In order for these equations to have a non-vanishing solution the determinant of the $1 + k + r$ equations must satisfy the well-known condition of being 0, i. e.,

$$\text{(j)} \qquad \begin{vmatrix} (S-a_{00}) & a_{01} \cdots a_{0k} & 0 \cdots 0 \\ (0-a_{10}) & a_{11} \cdots a_{1k} & c_{11} \cdots c_{r1} \\ \vdots & \vdots & \vdots \\ (0-a_{k0}) & a_{k1} \cdots a_{kk} & c_{1k} \cdots c_{rk} \\ (0-0) & c_{11} \cdots c_{1k} & 0 \cdots 0 \\ \vdots & \vdots & \vdots \\ (0-0) & c_{r1} \cdots c_{rk} & 0 \cdots 0 \end{vmatrix} = 0.$$

Treating the first column as a sum of 2 columns as indicated and employing the usual rule for expressing the determinant as the sum of two determinants, we find the minimum value of S to be given by

$$\text{(k)} \qquad S_{min} = \frac{\Delta}{\Delta_{00}},$$

where

$$\text{(l)} \qquad \Delta = \begin{vmatrix} a_{00} & a_{01} \cdots a_{0k} & 0 \cdots 0 \\ a_{10} & a_{11} \cdots a_{1k} & c_{11} \cdots c_{r1} \\ \vdots & \vdots & \vdots \\ a_{k0} & a_{k1} \cdots a_{kk} & c_{1k} \cdots c_{rk} \\ 0 & c_{11} \cdots c_{1k} & 0 \cdots 0 \\ \vdots & \vdots & \vdots \\ 0 & c_{r1} \cdots c_{rk} & 0 \cdots 0 \end{vmatrix}$$

and Δ_{00} is the minor of a_{00} in Δ.

It should be noted that the values of the a_p and λ_u which yield the extremum of F(or the values of the a_p which yield the minimum value of S) are given by the last $k + r$ linear equations in (i) with $a_0 = 1$.

CHAPTER IX

APPLICATIONS OF NORMAL REGRESSION THEORY TO ANALYSIS OF VARIANCE PROBLEMS

In this chapter we shall consider some applications of normal regression theory together with the general significance test embodied in Theorem (A), §8.3, to certain problems in the field of statistical analysis known as analysis of variance. This field of analysis is due primarily to R. A. Fisher.

9.1 Testing for the Equality of Means of Normal Populations with the Same Variance

Suppose $O_{n_p}(y_{p\alpha})$, $\alpha = 1,2,\ldots,n_p$, $p = 1,2,\ldots,k$, are samples from $N(a_1,\sigma^2)$, $N(a_2,\sigma^2)$, ..., $N(a_k,\sigma^2)$ respectively, and that it is desired to test the statistical hypothesis $H_0(a_1=a_2=\ldots=a_k)$ specified as follows:

$$\Omega: \quad -\infty < a_p < \infty, \qquad \sigma^2 > 0, \qquad p = 1,2,\ldots,k$$
$$\omega: \quad a_p = a, \quad -\infty < a < \infty, \qquad \sigma^2 > 0, \qquad p = 1,2,\ldots,k.$$

In other words H_0 is the statistical hypothesis that all of the samples are drawn from normal populations with identical means, given that the populations are normal and have equal variances. The probability element for the k samples is

$$\text{(a)} \qquad \left[\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{p=1}^{k}\sum_{\alpha=1}^{n_p}(y_{p\alpha}-a_p)^2}\Big] \prod_{p=1}^{k}\prod_{\alpha=1}^{n_p} dy_{p\alpha}, \qquad \left(n=\sum_{p=1}^{k} n_p\right).$$

Maximizing the likelihood function (i. e., the expression in [ ]) for variations of the parameters over Ω, we obtain

$$\text{(b)} \qquad \hat{\sigma}^2_{\Omega} = \frac{1}{n}\sum_{p=1}^{k}\sum_{\alpha=1}^{n_p}(y_{p\alpha}-\bar{y}_p)^2,$$

where $\bar{y}_p = \frac{1}{n_p}\sum_{\alpha=1}^{n_p} y_{p\alpha}$, the mean of the y's in the p-th sample. Maximizing the likelihood for variations of the parameters over ω we find

(c) $$\hat{\sigma}_{\omega}^2 = \frac{1}{n}\sum_{p=1}^{k}\sum_{\alpha=1}^{n_p}(y_{p\alpha}-\bar{y}\,)^2,$$

where $\bar{y} = \frac{1}{n}\sum_{p=1}^{k}\sum_{\alpha=1}^{n_p} y_{p\alpha} = \frac{1}{n}\sum_{p=1}^{k} n_p\bar{y}_p$. Now q_1 and q_2 of Theorem (A), §8.3, are as follows:

(d) $$q_1 = \frac{n\hat{\sigma}_{\Omega}^2}{\sigma^2}, \qquad q_2 = \frac{n(\hat{\sigma}_{\omega}^2 - \hat{\sigma}_{\Omega}^2)}{\sigma^2} = \frac{\sum_{p=1}^{k} n_p(\bar{y}_p-\bar{y})^2}{\sigma^2}.$$

Assuming $H(a_1=a_2=\ldots=a_k)$ is true (i. e., that $a_1 = a_2 = \ldots = a_k$) it follows from Theorem (A), §8.3, that q_1 and q_2 are independently distributed according to χ^2-laws with $n - k$ and $k - 1$ degrees of freedom respectively. Hence

(e) $$F = \frac{(n-k)q_2}{(k-1)q_1} = \frac{(n-k)\sum_{p=1}^{k} n_p(\bar{y}_p-\bar{y})^2}{(k-1)\sum_{p=1}^{k}\sum_{\alpha=1}^{n_p}(y_{p\alpha}-\bar{y}_p)^2}$$

is distributed according to $h_{k-1,n-k}(F)\,dF$.

To see exactly how this problem is an application of Theorem (A), the reader should refer to §8.41, Case 1. It will be noted that the set of k samples, 0_{n_1}, 0_{n_2},..., 0_{n_k}, can be regarded as a single sample of size n ($n = \sum_{p=1}^{k} n_p$) from a population with distribution $N(\sum_{p=1}^{k} a_p x_p, \sigma^2)$, where $x_1 = 1$, $x_2 = \ldots = x_k = 0$ for 0_{n_1}, $x_2 = 1$, $x_1 = x_3 = \ldots = x_k = 0$ for 0_{n_2}, and so on. The hypothesis $H(a_1=a_2=\ldots=a_k)$ is that all a_p are equal, i. e. $a_p = \sum_{q=1}^{k} c^{qp}a'_q$, where $c^{1p} = 1$, $a'_1 = a$, $a'_q = 0$, $q = 2,3,\ldots,k$.

9.2 Randomized Blocks or Two-way Layouts

Suppose y_{ij} ($i = 1,2,\ldots,r$, $j = 1,2,\ldots,s$) are random variables independently distributed according to $N(m+R_i+C_j,\sigma^2)$ where $\sum_{i=1}^{r} R_i = \sum_{j=1}^{s} C_j = 0$, and that we wish to test on basis of the y_{ij} the hypothesis $H[(C_j) = 0]$ specified as follows:

$$\Omega: \begin{cases} -\infty < m,\ R_i,\ C_j < \infty,\ \sigma^2 > 0, & i = 1,2,\ldots,r;\ j = 1,2,\ldots,s \\ \sum_{i=1}^{r} R_i = \sum_{j=1}^{s} C_j = 0 \end{cases}$$

$$\omega: \left\{ \text{The subspace in } \Omega \text{ obtained by setting each } C_j = 0. \right.$$

The ω space is simply the subspace in Ω for which the C_j are all 0. The probability element for the sample (i. e. the y_{ij}) is

(a) $$\left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{rs} e^{-\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij}-m-R_i-C_j)^2}\right]\prod_{i,j} dy_{ij}.$$

The sum of squares in the exponent of (a) may be written as

(b)
$$S = \sum_{i,j}[(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}) + (\bar{y}_{i\cdot}-\bar{y}-R_i) + (\bar{y}_{\cdot j}-\bar{y}-C_j) + (\bar{y}-m)]^2,$$
$$= \sum_{i,j}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})^2 + \sum_{i,j}(\bar{y}_{i\cdot}-\bar{y}-R_i)^2 + \sum_{i,j}(\bar{y}_{\cdot j}-\bar{y}-C_j)^2 + rs(\bar{y}-m)^2,$$

where $\bar{y} = \frac{1}{rs}\sum_{i,j} y_{ij}$, $\bar{y}_{i\cdot} = \frac{1}{s}\sum_{j} y_{ij}$ $\bar{y}_{\cdot j} = \frac{1}{r}\sum_{i} y_{ij}$. Maximizing the likelihood function in [ ] (which is equivalent to minimizing S as far as m, the R_i and C_j are concerned) for variations of the parameters over Ω, we find

$$\hat{m} = \bar{y},\ \hat{R}_i = \bar{y}_{i\cdot} - \bar{y},\ \hat{C}_j = \bar{y}_{\cdot j} - \bar{y},$$
$$\hat{\sigma}^2_{\Omega} = \frac{1}{rs}\sum_{i,j}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})^2.$$

Maximizing the likelihood function for variations of the parameters over ω, we simply set each of the C_j equal to zero and maximize for variations of σ^2, m, R_i (subject to $\sum_{1}^{r} R_i = 0$). We find

$$\hat{m} = \bar{y}, \qquad \hat{R}_i = \bar{y}_{i\cdot} - \bar{y}$$
$$\hat{\sigma}^2_{\omega} = \frac{1}{rs}\sum_{i,j}(y_{ij}-\bar{y}_{i\cdot})^2.$$

It may be readily verified that

$$\hat{\sigma}^2_{\omega} = \hat{\sigma}^2_{\Omega} + \frac{1}{rs}\sum_{i,j}(\bar{y}_{\cdot j}-\bar{y})^2.$$

q_1 and q_2 of Theorem (A), §8.3, are given by

$$q_1 = \frac{rs\hat{\sigma}^2_{\Omega}}{\sigma^2}, \quad q_2 = \frac{rs(\hat{\sigma}^2_{\omega}-\hat{\sigma}^2_{\Omega})}{\sigma^2} = \frac{\sum_{i,j}(\bar{y}_{\cdot j}-\bar{y})^2}{\sigma^2}.$$

It follows from Theorem (A), §8.3, (See Case 3, §8.43) that q_1 and q_2 are independently distributed according to χ^2-laws with (r-1)(s-1) and (s-1) degrees of freedom respectively, when $H[(C_j) = 0]$ is true. Hence, under the same conditions,

$$F = \frac{(r-1)q_2}{q_1} = \frac{(r-1)\sum_{i,j}(\bar{y}_{\cdot j}-\bar{y})^2}{\sum_{i,j}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})^2}$$

is distributed according to $h_{(s-1),(r-1)(s-1)}(F)dF$, and is equivalent to the likelihood ratio criterion for testing $H[(C_j) = 0]$ using the upper tail of the distribution for obtaining critical values of F for given significance levels.

In an entirely similar manner we may derive an F test for the hypothesis

$H[(R_i)=0]$ defined as follows:

Ω: { Same as for Ω in definition of $H[(C_j)=0]$

ω: { The subspace in Ω obtained by setting each $R_i = 0$.

Following steps similar to those followed for $H[(C_j)=0]$ we find for $H[(R_i)=0]$,

$$(c) \qquad F = \frac{(s-1)\sum_{i,j}(\bar{y}_{i\cdot}-\bar{y})^2}{\sum_{i,j}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})^2} \,,.$$

which will be distributed according to $h_{(r-1),(r-1)(s-1)}(F)dF$, when $H[(R_i)=0]$ is true.

The applicability of Theorem (A), §8.3, in testing $H[(C_j)=0]$ is evident when it is noted that under Ω the y_{ij} can be regarded as a sample of size rs from a population having a distribution of the form $N(\sum_{p=1}^{r+s+1} a_p x_p, \sigma^2)$ in which there are two homogeneous linear conditions on the a_p (the a_p being written in place of the m, R_i, C_j and each x having the value 0 or 1) whereas under ω there would be s + 1 linear conditions on the a_p (or s - 1 linear conditions in addition to those already imposed under Ω). Both $H[(C_j)=0]$ and $H[(R_i)=0]$ come under Case 3, §8.43.

If the y_{ij}, $i = 1,2,\ldots,r$; $j = 1,2,\ldots,s$, are considered in a rectangular array with i referring to rows and j to columns, then it will be seen that we are assuming that y_{ij} is a normally distributed random variable with a mean which is the sum of three parts: a general constant m, a specific constant R_i associated with the i-th row and a specific constant C_j associated with the j-th column (where $\sum_i R_i = \sum_j C_j = 0$). The variance is assumed to be independent of i and j. Statistically speaking R_i is often referred to as *effect (or main effect) due to the i-th row*, and C_j the *effect (or main effect) due to the j-th column*. $H[(R_i)=0]$ is therefore the hypothesis that row effects are zero no matter what the values of m and column effects. The quantity $\sum_{i,j}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})^2$ is often referred to as "error" or "residual" sum of squares after row and column effects are removed, and when divided by (r-1) (s-1) the resulting expression provides an unbiased estimate of σ^2 no matter what the values of m, the R_i and C_j. $\sum_{ij}(\bar{y}_{i\cdot}-y)^2$ is usually referred to as *sum of squares due to rows*, and when divided by r - 1, the resulting quotient provides an unbiased estimate of σ^2 (and, as we have seen, independent of that obtained by using "error" sum of squares) if the $R_i = 0$, no matter what values the C_j and m may have. A similar statement holds for $\sum_{i,j}(\bar{y}_{\cdot j}-\bar{y})^2$. It can be shown by Cochran's Theorem and by the use of moment generating functions, although we shall not do so here, that $\frac{1}{\sigma^2}\sum_{i,j}(y_{ij}-y_{i\cdot}-y_{\cdot j}+y)^2$, $\frac{1}{\sigma^2}\sum_{i,j}(y_{i\cdot}-\bar{y})^2$, $\frac{1}{\sigma^2}\sum_{i,j}(\bar{y}_{\cdot j}-\bar{y})^2$ are independently distributed according to χ^2-laws with (r-1) (s-1), (r-1), (s-1) degrees of freedom respectively, if the R_i and

C_j are all zero, and furthermore the sum of the three quantities is $\frac{1}{\sigma^2}\sum_{i,j}(y_{ij}-\bar{y})^2$, which is distributed according to the χ^2-law with rs-1 degrees of freedom if each R_i and each C_j is zero.

These various sums of squares together with their degrees of freedom are commonly set forth in an analysis of variance table as follows:

Variation Due to	Sum of Squares	Degrees of Freedom
Rows	$S_R = \sum_{i,j}(\bar{y}_{i.}-\bar{y})^2$	r - 1
Columns	$S_C = \sum_{i,j}(\bar{y}_{.j}-\bar{y})^2$	s - 1
Error	$S_E = \sum_{i,j}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})^2$	(r-1) (s-1)
Total	$S = \sum_{i,j}(y_{ij}-\bar{y})^2$	rs - 1

The main facts regarding the constituents of this Table may be summarized as follows:

(1) $S = S_R + S_C + S_E$.

(2) S_R/σ^2, S_C/σ^2, S_E/σ^2 are independently distributed according to χ^2-laws with (r-1), (s-1), (r-1)(s-1) degrees of freedom respectively if all R_i and C_j are zero.

(3) $F = \frac{(s-1)S_R}{S_E}$ is distributed according to $h_{(r-1),(r-1)(s-1)}(F)dF$ when $H[(R_i)=0]$ is true.

(4) $F = \frac{(r-1)S_C}{S_E}$ is distributed according to $h_{(s-1),(r-1)(s-1)}(F)dF$ when $H[(C_j)=0]$ is true.

(5) S_E/σ^2 is distributed according to the χ^2-law with (r-1)(s-1) degrees of freedom for any parameter point in Ω (i. e. no matter what values the R_i and C_j may have).

(6) S/σ^2 is distributed according to the χ^2-law with rs-1 degrees of freedom if all R_i and C_j are zero.

The theory discussed in this section has been widely used in what are called designed experiments, particularly in agricultural science. For example, rows in our rectangular array may be associated with r different varieties of wheat, columns with s different types of fertilizer, and y_{ij} with the yield of wheat on the plot of soil associated with the i-th variety and j-th fertilizer, it being assumed that plots are of the same size and the soil homogeneous for all plots. In such an application, we emphasize that the fundamental assumptions are that the yield on the plot associated with the i-th

variety and the j-th fertilizer may be regarded as a normally distributed random variable having mean value of the form $m + R_i + C_j$ (where $\sum_i R_i = \sum_j C_j = 0$), and a variance σ^2 which has the same value for all i and j. The question of whether the assumptions are tenable in any given case is one for the individual applying the method to settle. In this example $H[(C_j)=0]$ would be the hypothesis that fertilizer effects on yield are all equal no matter what the variety effects may be.

9.3 Three-way and Higher Order Layouts; Interaction

The analysis presented in §9.2 can be extended to three-way and higher order layouts. In this section we shall consider in detail the three-way layout. Let y_{ijk} ($i = 1,2,\ldots,r$; $j = 1,2,\ldots,s$; $k = 1,2,\ldots,t$) be random variables distributed independently according to

(a) $$N(m+\bar{I}_{ijk},\sigma^2),$$

where

(b) $$\bar{I}_{ijk} = I_{ijo} + I_{iok} + I_{ojk} + I_{ioo} + I_{ojo} + I_{ook},$$

where each set of I's on the right hand side of (b) is such that when summed over each index the sum is zero. Thus there are $(r-1)(s-1)$ linearly independent constants in the set $\{I_{ijo}\}$, $(r-1)$ such constants in the set $\{I_{ioo}\}$, with similar statements holding for the remaining sets. For convenience, we may consider y_{ijk} as a random variable associated with the cell in the i-th row, j-th column and k-th layer of a three-dimensional rectangular array of cells. The mean value of y_{ijk} is given in (a), in which the I_{ioo}, the I_{ojo} and the I_{ook} are row, column and layer *main effects*, respectively; the I_{ijo} are *row-column interactions**, the I_{iok} *row-layer interactions**, and the I_{ojk} are *column-layer interactions**.

The probability element of the y_{ijk} is

(c) $$\prod_{i,j,k} N(m+\bar{I}_{ijk},\sigma^2)dy_{ijk}.$$

The sum of squares in the exponent of (c) is

(d) $$S = \sum_{i,j,k}(y_{ijk}-\bar{I}_{ijk}-m)^2.$$

Now let

(e)
$$\bar{y} = \frac{1}{rst}\sum_{i,j,k} y_{ijk},$$
$$\bar{y}_{i..} = \frac{1}{st}\sum_{j,k} y_{ijk}, \text{ with similar meanings for } \bar{y}_{.j.} \text{ and } \bar{y}_{..k},$$
$$\bar{y}_{ij.} = \frac{1}{t}\sum_{k} y_{ijk}, \text{ with similar meanings for } \bar{y}_{i.k} \text{ and } \bar{y}_{.jk},$$
$$\bar{Y}_{i..} = \bar{y}_{i..}-\bar{y}, \text{ with similar meanings for } \bar{Y}_{.j.} \text{ and } \bar{Y}_{..k},$$

*These are called *first-order* interactions.

$\bar{Y}_{ij\cdot} = \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}$, with similar meanings for $\bar{Y}_{i\cdot k}$ and $\bar{Y}_{\cdot jk}$,

(e)
$$S_{\cdots} = \sum_{i,j,k} (y_{ijk} - \bar{y}_{ij\cdot} - \bar{y}_{i\cdot k} - \bar{y}_{\cdot jk} + \bar{y}_{i\cdot\cdot} + \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot k} - \bar{y})^2$$

$$= \sum_{i,j,k} (y_{ijk} - \bar{y}_{ij\cdot} - \bar{y}_{i\cdot k} - \bar{y}_{\cdot jk} + \bar{y}_{i\cdot\cdot} + \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot k} - \bar{y})^2,$$

$$S_{\cdot\cdot o} = \sum_{i,j,k} (\bar{Y}_{ij\cdot} - I_{ijo})^2, \text{ with similar meanings for } S_{\cdot o\cdot} \text{ and } S_{o\cdot\cdot},$$

$$S_{\cdot oo} = \sum_{i,j,k} (\bar{Y}_{i\cdot\cdot} - I_{ioo})^2, \text{ with similar meanings for } S_{o\cdot o} \text{ and } S_{oo\cdot},$$

$$S_{ooo} = \sum_{i,j,k} (\bar{y} - m)^2.$$

Let $S^{o}_{\cdot\cdot o}$ be the value of $S_{\cdot\cdot o}$ with each $I_{ijo} = 0$, with similar meanings for $S^{o}_{\cdot o\cdot}$, $S^{o}_{o\cdot\cdot}$, $S^{o}_{\cdot oo}$, $S^{o}_{o\cdot o}$, $S^{o}_{oo\cdot}$.

We may write

(f)
$$S = \sum_{i,j,k} [(y_{ijk} - \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot k} - \bar{Y}_{\cdot jk} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot k} - \bar{y})$$
$$+ (\bar{Y}_{ij\cdot} - I_{ijo}) + (\bar{Y}_{i\cdot k} - I_{iok}) + (\bar{Y}_{\cdot jk} - I_{ojk})$$
$$+ (\bar{Y}_{i\cdot\cdot} - I_{ioo}) + (\bar{Y}_{\cdot j\cdot} - I_{ojo}) + (\bar{Y}_{\cdot\cdot k} - I_{ook}) + (\bar{y} - m)]^2.$$

Squaring the quantity in [ ], keeping the expressions within the parentheses intact, and summing with respect to i, j, k, we obtain

(g)
$$S = S_{\cdots} + S_{\cdot\cdot o} + S_{\cdot o\cdot} + S_{o\cdot\cdot} + S_{\cdot oo} + S_{o\cdot o} + S_{oo\cdot} + S_{ooo}.$$

It follows from Cochran's Theorem, §5.24 , (and can also be shown by moment-generating functions) that the eight sums of squares on the right side of (g), each divided by σ^2, are independently distributed according to χ^2-laws with $(r-1)(s-1)(t-1)$, $(r-1)(s-1)$, $(r-1)(t-1)$, $(s-1)(t-1)$, $(r-1)$, $(s-1)$, $(t-1)$, 1 degrees of freedom, respectively, if the y_{ijk} are distributed according to (a).

The sums of squares in (g) provide the basis for testing various hypotheses concerning the interactions I_{ijo}, I_{iok}, I_{ojk} and the main effects I_{ioo}, I_{ojo}, I_{ook}. For example, suppose we wish to test the hypothesis that row-column interaction is zero (i. e. each $I_{ijo}=0$) no matter what the row-layer and column-layer interactions and main effects may be. This hypothesis, say $H[(I_{ijo})=0]$, may be specified as follows:

(h)
$$\Omega: \begin{cases} -\infty < m,\ I_{ijo},\ I_{ioj},\ I_{ojk},\ I_{ioo},\ I_{ojo},\ I_{ook} < \infty,\ \sigma^2 > 0, \\ \text{for all } i,j,k, \text{ the sum of the I's in each set over any} \\ \text{index being 0.} \end{cases}$$

$$\omega: \quad \{ \text{ Subspace of } \Omega \text{ obtained by setting each } I_{ijo} = 0.$$

Maximizing the likelihood in (c) for variations of the parameters over Ω, we find

$$\hat{\sigma}^2_{\Omega} = \frac{1}{rst} S_{...}, \tag{i}$$

and maximizing the likelihood for variations of the parameters over ω, we find

$$\hat{\sigma}^2_{\omega} = \frac{1}{rst}(S_{...} + S^{o}_{..o}). \tag{j}$$

It should be noted that in maximizing the likelihood over Ω we obtain as maximum likelihood estimates of I_{ijo}, I_{iok}, I_{ojk}, I_{ioo}, I_{ojo}, I_{ook} the quantities $\bar{Y}_{ij.}$, $\bar{Y}_{i.k}$, $\bar{Y}_{.jk}$, $\bar{Y}_{i..}$, $\bar{Y}_{.j.}$, $\bar{Y}_{..k}$, respectively.

When the hypothesis $H[(I_{ijo})=0]$ is true it follows from Theorem (A), §8.3 (see Case 3, §8.43), that

$$q_1 = \frac{rst\hat{\sigma}^2_{\Omega}}{\sigma^2} = \frac{S_{...}}{\sigma^2}, \qquad q_2 = \frac{rst(\hat{\sigma}^2_{\omega} - \hat{\sigma}^2_{\Omega})}{\sigma^2} = \frac{S^{o}_{..o}}{\sigma^2} \tag{k}$$

are independently distributed according to χ^2-laws with $(r-1)(s-1)(t-1)$ and $(r-1)(s-1)$ degrees of freedom, respectively. Hence the F-ratio for testing this hypothesis is

$$\frac{(t-1)S^{o}_{..o}}{S_{...}},$$

which is distributed according to $h_{(r-1)(s-1),\,(r-1)(s-1)(t-1)}(F)dF$ when $H[(I_{ijo})=0]$ is true. In a similar manner F-ratios can be set up for testing the hypothesis of zero row-layer or zero column-layer interaction.

The constituents in (g) also provide a method of testing the hypothesis of no interaction between rows and columns in a two-way layout from t ($t \geq 2$) replications of the layout. This hypothesis amounts to the hypothesis that effects due to rows and columns are additive on the mean value of the y_{ij}, in which case the mean value of y_{ij} is of the form $m + I_{ioo} + I_{ojo}$. In this problem we consider y_{ijk} ($i = 1,2,\ldots,r$; $j = 1,2,\ldots,s$) as the variables associated with the k-th replicate, and assume the mean value of y_{ijk} to be $m + I_{ijo} + I_{ioo} + I_{ojo}$. The problem is to test the hypothesis that each $I_{ijo} = 0$. This hypothesis which will be called $H'[(I_{ijo})=0]$ is specified as follows:

(l)

Ω: $\begin{cases} -\infty < m,\ I_{ijo},\ I_{ioo},\ I_{ojo} < \infty,\ \sigma^2 > 0, \text{ for each } i \text{ and} \\ j, \text{ where the sum of the I's in each set over each index} \\ \text{is } 0. \end{cases}$

ω: [The subspace of Ω obtained by setting each $I_{ijo} = 0$.

Maximizing the likelihood function in (c) for variations of the parameters over Ω, we find

(m) $$\hat{\sigma}^2_{\Omega} = \frac{1}{rst}\,(S_{...}+S^{o}_{.o.}+S^{o}_{o..}+S^{o}_{oo.}),$$

and similarly

(n) $$\hat{\sigma}^2_{\omega} = \frac{1}{rst}\,(S_{...}+S^{o}_{..o}+S^{o}_{.o.}+S^{o}_{o..}+S^{o}_{oo.}).$$

By Theorem (A), §8.3, it follows that

(o) $$q_1 = \frac{rst\hat{\sigma}^2_{\Omega}}{\sigma^2} = \frac{S_{...}+S^{o}_{.o.}+S^{o}_{o..}+S^{o}_{oo.}}{\sigma^2},$$

$$q_2 = \frac{rst(\hat{\sigma}^2_{\omega}-\hat{\sigma}^2_{\Omega})}{\sigma^2} = \frac{S^{o}_{..o}}{\sigma^2}$$

are independently distributed according to χ^2-laws with $rs(t-1)$ and $(r-1)(s-1)$ degrees of freedom, respectively, when $H'[(I_{1jo})=0]$, is true, and hence under the same assumptions

(p) $$F = \frac{rs(t-1)\,S^{o}_{..o}}{(r-1)(s-1)(S_{...}+S^{o}_{.o.}+S^{o}_{o..}+S^{o}_{oo.})}$$

is distributed according to

$$h_{(r-1)(s-1),\ rs(t-1)}(F)dF.$$

In a similar manner, the existence of second-order interaction in a three-way layout may be tested on basis of replications of the three-way layout. This problem, however, leads us into four-way layouts and the details must be left to the reader.

Suppose we are interested in testing the hypothesis $H[(I_{ioo})=0]$ that the $I_{ioo} = 0$ no matter what the interactions and main effects due to columns and layers may be. This hypothesis may be specified as follows:

(q) Ω: | Same as Ω in (h).

ω: | Subspace of Ω for which each $I_{ioo} = 0$.

We have

$$\hat{\sigma}^2_{\Omega} = \frac{1}{rst}\,S_{...},$$

and

$$\hat{\sigma}^2_{\omega} = \frac{1}{rst}\,(S_{...} + S^{o}_{.oo}),$$

and hence by Theorem (A), §8.3 ,

(r)
$$q_1 = \frac{rst}{\sigma^2}\hat{\sigma}_\Omega^2 = \frac{S_{...}}{\sigma^2},$$
$$q_2 = \frac{rst}{\sigma^2}(\hat{\sigma}_\omega^2 - \hat{\sigma}_\Omega^2) = \frac{S^0_{\cdot 00}}{\sigma^2}$$

are independently distributed according to χ^2-laws with $(r-1)(s-1)(t-1)$ and $(r-1)$ degrees of freedom, respectively, when $H[(I_{i00})=0]$, is true, and the F-ratio for the hypothesis is

$$\frac{(s-1)(t-1)S^0_{\cdot 00}}{S_{...}},$$

which is distributed according to $h_{(r-1),(r-1)(s-1)(t-1)}(F)dF$. Similar tests exist for testing the hypothesis that the $I_{0j0} = 0$ or that the $I_{00k} = 0$.

Suppose the interactions I_{ij0}, I_{0jk}, and I_{i0k} are all zero and that it is desired to test the hypothesis that the main effects due to rows are 0, i. e., $I_{i00} = 0$. This hypothesis say $H'[(I_{i00})=0]$ may be specified as follows:

(s)
$$\Omega: \begin{cases} -\infty < m, I_{i00}, I_{0j0}, I_{00k} < \infty, \sigma^2 > 0, \\ \sum_i I_{i00} = \sum_j I_{0j0} = \sum_k I_{00k} = 0. \end{cases}$$

ω: { Subspace of Ω obtained by setting each $I_{i00} = 0$.

We find

$$\hat{\sigma}_\Omega^2 = \frac{1}{rst}(S_{...} + S^0_{\cdot\cdot 0} + S^0_{\cdot 0\cdot} + S^0_{0\cdot\cdot}),$$

$$\hat{\sigma}_\omega^2 = \frac{1}{rst}(S_{...} + S^0_{\cdot\cdot 0} + S^0_{\cdot 0\cdot} + S^0_{0\cdot\cdot} + S^0_{\cdot 00}),$$

and hence

$$q_1 = \frac{rst\hat{\sigma}_\Omega^2}{\sigma^2} = \frac{S_{...} + S^0_{\cdot\cdot 0} + S^0_{\cdot 0\cdot} + S^0_{0\cdot\cdot}}{\sigma^2},$$

$$q_2 = \frac{rst(\hat{\sigma}_\omega^2 - \hat{\sigma}_\Omega^2)}{\sigma^2} = \frac{S^0_{\cdot 00}}{\sigma^2},$$

which are distributed independently according to χ^2-laws with $rst - r - s - t + 2$ and $r - 1$ degrees of freedom, respectively, when $H'[(I_{i00})=0]$ is true. The F-ratio is

$$\frac{(rst - r - s - t + 2)\, S^0_{\cdot 00}}{(r-1)(S_{...} + S^0_{\cdot\cdot 0} + S^0_{\cdot 0\cdot} + S^0_{0\cdot\cdot})},$$

which has the distribution $h_{(r-1),\,(rst-r-s-t+2)}(F)dF$, when $H'[(I_{i00})=0]$ is true.

The difference between the F-ratio for testing $H[(I_{i00})=0]$ and that for testing $H'[(I_{i00})=0]$ should be noted. In the first hypothesis the interactions are not assumed to be

different from zero, and in the second one the interactions are assumed to be zero. The sum of squares in the denominator of F for the first hypothesis is simply $S_{...}$ while it is $S_{...} + S^{o}_{..o} + S^{o}_{.o.} + S^{o}_{o..}$ in the F for the second hypothesis. The terms $S^{o}_{..o}$, $S^{o}_{.o.}$, $S^{o}_{o..}$, are commonly known as interaction sums of squares, and the process of adding these to the error sum of squares $S_{...}$ in the case of testing $H'[(I_{ioo})=0]$ is often referred to as confounding first-order interactions with error. Of course, the hypothesis may be set up in such a way that only two (or even only one) of the interaction sum of squares will be confounded with error. The term confounding as it is commonly used is more general than it is in the sense used above. For example, if layer effects (I_{ook}) are assumed to be zero throughout the hypothesis specified by (s) we would have found not only all first-order interaction sum of squares but also layer effect sum of squares $S^{o}_{oo.}$ confounded with $S_{...}$.

There are many hypotheses which can be tested on basis of the S's on the right hand side of (g), and we shall make no attempt to catalogue them here. It is perhaps sufficient to summarize the constituents of the various possible tests in the following analysis of variance table (the $\sum$ extending over all values of i, j, k in each case):

Variation Due To	Sum of Squares	Degrees of Freedom
Rows	$S^{o}_{.oo} = \sum(\bar{y}_{i..} - \bar{y})^2$	r - 1
Columns	$S^{o}_{o.o} = \sum(\bar{y}_{.j.} - \bar{y})^2$	s - 1
Layers	$S^{o}_{oo.} = \sum(\bar{y}_{..k} - \bar{y})^2$	t - 1
Row-Column Interaction	$S^{o}_{..o} = \sum(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y})^2$	(r-1)(s-1)
Row-Layer Interaction	$S^{o}_{.o.} = \sum(\bar{y}_{i.k} - \bar{y}_{i..} - \bar{y}_{..k} + \bar{y})^2$	(r-1)(t-1)
Column-Layer Interaction	$S^{o}_{o..} = \sum(\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y})^2$	(s-1)(t-1)
Error	$S^{o}_{...} = \sum(y_{ijk} - \bar{y}_{ij.} - \bar{y}_{i.k} - \bar{y}_{.jk} + \bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{..k} - \bar{y})^2$	(r-1)(s-1)(t-1)
Total	$S_T = \sum(y_{ijk} - \bar{y})^2$	rst - 1

9.4 Latin Squares

Suppose y_{ij} (i,j = 1,2,...,r) are random variables distributed according to $N(m + R_i + C_j + T_t, \sigma^2)$, where $\sum_{i=1}^{r} R_i = \sum_{j=1}^{r} C_j = \sum_{t=1}^{r} T_t = 0$, such that each T_t occurs in conjunction with each R_i once and only once, and with each C_j once and only once, each R_i occurring once and only once in conjunction with each C_j. Such an arrangement of combinations of attributes is known as a Latin Square arrangement. For a given r there are many

such arrangements, each of which can be represented in a square array in which the R_i would be *row effects*, the C_j *column effects* and the T_t *treatment effects*. For example, when r = 4, the following is a Latin Square arrangement of row, column and treatment effects:

$R_1+C_1+T_1$	$R_1+C_2+T_2$	$R_1+C_3+T_3$	$R_1+C_4+T_4$
$R_2+C_1+T_4$	$R_2+C_2+T_1$	$R_2+C_3+T_2$	$R_2+C_4+T_3$
$R_3+C_1+T_3$	$R_3+C_2+T_4$	$R_3+C_3+T_1$	$R_3+C_4+T_2$
$R_4+C_1+T_2$	$R_4+C_2+T_3$	$R_4+C_3+T_4$	$R_4+C_4+T_1$

Fisher and Yates (*Statistical Tables*, Oliver and Boyd, Edinburgh, 1938) have tabulated Latin Squares up to size 12 by 12.

Now consider the following hypothesis, say $H[(T_t)=0]$, to be tested on basis of the sample y_{ij}

$$\Omega: \begin{cases} -\infty < m, R_i, C_j, T_t < +\infty, \ \sigma^2 > 0, \\ \sum_1^r R_i = \sum_1^r C_j = \sum_1^r T_t = 0. \end{cases}$$

ω: {Subspace in Ω obtained by setting each $T_t = 0$.

In other words, we wish to test the hypothesis that the T_t are all zero, assuming that th y_{ij} are distributed according to $N(m+R_i+C_j+T_t, \sigma^2)$. The probability element of the y_{ij} is

$$\text{(a)} \qquad \left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{r^2} e^{-\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij}-m-R_i-C_j-T_t)^2}\right] \prod_{i,j} dy_{ij}.$$

The sum of squares S in the exponent may be written as

$$S = S_E + S_R + S_C + S_T + r^2(\bar{y}-m)^2,$$

where

$$S_E = \sum_{i,j}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}-\bar{y}_{(t)}+2\bar{y})^2,$$
$$S_R = \sum_{i,j}(\bar{y}_{i\cdot}-\bar{y}-R_i)^2,$$
$$S_C = \sum_{i,j}(\bar{y}_{\cdot j}-\bar{y}-C_j)^2,$$
$$S_T = \sum_{i,j}(\bar{y}_{(t)}-\bar{y}-T_t)^2,$$

where $\bar{y} = \frac{1}{r^2}\sum_{i,j} y_{ij}$, $\bar{y}_{i.} = \frac{1}{r}\sum_{j} y_{ij}$, $\bar{y}_{.j} = \frac{1}{r}\sum_{i} y_{ij}$ and $\bar{y}_{(t)} = \frac{1}{r}\sum_{(t)} y_{ij}$, $\sum_{(t)}$ denoting summation over all cells (i and j) in the Latin Square array in which T_t occurs. Let S_R^o be the value of S_R when the $R_i = 0$, with similar meanings for S_C^o and S_T^o.

MAXIMIZING Maximizing the likelihood function in (a) for variations of the parameters over Ω we find

$$\hat{m} = \bar{y}, \quad \hat{R}_i = \bar{y}_{i.} - \bar{y}, \quad \hat{C}_j = \bar{y}_{.j} - \bar{y}, \quad \hat{T}_t = \bar{y}_{(t)} - \bar{y}.$$

(b)
$$\hat{\sigma}_\Omega^2 = \frac{1}{r^2} S_E$$
$$= \frac{1}{r^2}\sum_{i,j}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}-\bar{y}_{(t)}+2\bar{y})^2.$$

Maximizing the likelihood for variations of the parameters over ω, we set $T_t = 0$ $(t = 1,2,\ldots,r)$ and maximize for variations of m, R_i, C_j, σ^2 $(\sum_i R_i = \sum_j C_j = 0)$. We find $\hat{m}$, $\hat{R}_i$, $\hat{C}_j$ to be the same as those obtained by maximizing over Ω, and

(c)
$$\hat{\sigma}_\omega^2 = \frac{1}{r^2}(S_E + S_T^o),$$
$$= \frac{1}{r^2}\sum_{i,j}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})^2.$$

It follows from Theorem (A), §8.3, (see Case 3, §8.43) that q_1 and q_2 are independently distributed according to the χ^2-laws with $(r-1)(r-2)$ and $(r-1)$ degrees of freedom respectively when $H[(T_t)=0]$ is true, where

$$q_1 = \frac{r^2\hat{\sigma}_\Omega^2}{\sigma^2},$$
$$q_2 = \frac{r^2(\hat{\sigma}_\omega^2 - \hat{\sigma}_\Omega^2)}{\sigma^2} = \frac{S_T^o}{\sigma^2},$$

and hence

(d)
$$F = \frac{(r-2)q_2}{q_1} = \frac{(r-2)S_T^o}{S_E}$$

is distributed according to $h_{(r-1)(r-1)(r-2)}(F)dF$ and is equivalent to the likelihood ratio ratio criterion for testing $H[(T_t)=0]$, it being understood, of course, that critical values of F for a given significance level are obtained by using the upper tail of the F distribution.

In a similar manner, if $H[(R_i)=0]$ denotes the hypothesis for which Ω is identical with that for $H[(T_t)=0]$ while ω is the subspace in Ω for which $R_1 = R_2 = \ldots = R_r = 0$,

then we obtain for F the following:

(e) $$F = \frac{(r-2)S_R^o}{S_E},$$

which is distributed according to $h_{(r-1),(r-2)(r-1)}(F)dF$ when $H[(R_i)=0]$ is true.

An entirely similar hypothesis, say $H[(C_j)=0]$ may be defined by considering ω as the subspace in Ω for which $C_1 = C_2 = \ldots = C_r = 0$, and an F similar to (e) is obtained with the same distribution as that of F defined by (e).

We may summarize in the following analysis of variance table

Variation Due to	Sum of Squares	Degrees of Freedom
Rows	$S_R^o = r\sum_i (\bar{y}_{i.}-\bar{y})^2$	r - 1
Columns	$S_C^o = r\sum_j (\bar{y}_{.j}-\bar{y})^2$	r - 1
Treatments	$S_T^o = r\sum_{(t)} (\bar{y}_{(t)}-\bar{y})^2$	r - 1
Error	$S_E = \sum_{i,j} (y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}-\bar{y}_{(t)}+2\bar{y})^2$	(r-1)(r-2)
Total	$S = \sum_{i,j} (y_{ij}-\bar{y})^2$	r^2 - 1

The main properties relating to the constituents of this table are the following:

(1) $S = S_R^o + S_C^o + S_T^o + S_E$.

(2) S_R^o/σ^2, S_C^o/σ^2, S_T^o/σ^2, S_E/σ^2 are independently distributed according to χ^2-laws with r-1, r-1, r-1, (r-1)(r-2) degrees of freedom respectively, when all R_i, C_j and T_t are zero.

(3) $F = \frac{(r-1)S_R^o}{S_E}$ is distributed according to $h_{(r-1),(r-1)(r-2)}(F)dF$ when $H[(R_i)=0]$ is true.

(4) $F = \frac{(r-1)S_C^o}{S_E}$ is distributed according to $h_{(r-1),(r-1)(r-2)}(F)dF$ when $H[(C_j)=0]$ is true.

(5) $F = \frac{(r-1)S_T^o}{S_E}$ is distributed according to $h_{(r-1),(r-1)(r-2)}(F)dF$ when $H[(T_t)=0]$ is true.

(6) S_E/σ^2 is distributed according to the χ^2-law with (r-1)(r-2) degrees of freedom for any parameter point in Ω (i. e. no matter what values m and the R_i, C_j, T_t may have).

(7) S/σ^2 is distributed according to the χ^2-law with r^2-1 degrees of freedom when all R_i, C_j and T_t are zero.

The reader will find it instructive to verify that S_E is the minimum of $\sum_{i,j}(y_{ij}-m-R_i-C_j-T_t)^2$ for variations of the m, R_i, C_j and T_t subject to the restrictions $\sum_1^r R_i = \sum_1^r C_j = \sum_1^r T_t = 0$, and can be obtained by applying formula (k) of §8.5, noting that all $x_{p\alpha}$ are 0 or 1.

As in the case of two- and three-way and higher order layouts, Latin Square lay-outs have been widely used in agricultural experiments. For example in studying the effects of r types of fertilizer on yields of a certain variety of wheat, it is common to lay out a square array of r^2 plots of equal area and to associate row and column effects with variations in fertility of soil and associate treatments with different fertilizers. The main assumption in such an application is that variation in fertility of soil from plot to plot is such that yield on the plot in the i-th row and j-th column may be regarded as a normally distributed random variable y_{ij} with mean value of the form $m + R_i + C_j + T_t$, (where $\sum_i R_i = \sum_j C_j = \sum_t T_t = 0$, T_t being the effect of the t-th treatment) and variance σ^2 which is the same for all plots.

Latin Square lay-outs have also been tried out in other fields, for example in industrial research.

9.5 Graeco-Latin Squares

Higher order Latin Squares, known as Graeco-Latin Squares may be treated in much the same manner as Latin Squares. A Graeco-Latin square involving, for example, a four-way classification may be defined as follows: Let $|\alpha_i|$, $|\beta_i|$, $|\gamma_i|$, $|\delta_i|$, $i = 1,2,\ldots,r$, be four sets of mutually exclusive attributes. Let r^2 objects be arranged in such a way that r of the objects have attribute α_i, r have attribute β_i, r have attribute γ_i, and r have attribute δ_i, $i = 1,2,\ldots,r$, and in such a way that exactly one object has the combination of attributes (α_i, β_j), $i,j=1,2,\ldots,r$, exactly one has the combination (α_i,γ_j), and so on for each of the combinations (α_i,δ_j), (β_i,γ_j), (β_i,δ_j), (γ_i,δ_j). We may conveniently allow the α_i to refer to rows, β_i to columns, γ_i to treatments in an ordinary Latin Square and let δ_i refer to the fourth classification. Let y_{ij} $(i,j = 1,2,\ldots,r)$ be random variables distributed according to $N(m+R_i+C_j+T_t+U_u,\sigma^2)$ where R_i, C_j, T_t, U_u are *effects due to* α_i, β_j, γ_t, δ_u, and where $\sum_1^r R_i = \sum_1^r C_j = \sum_1^r T_t = \sum_1^r U_u = 0$. As a matter of fact, we may consider the four-way classification Graeco-Latin square as a superposition of the two Latin squares $|\alpha_i|$, $|\beta_i|$, $|\gamma_i|$ and $|\alpha_i|$, $|\beta_i|$, $|\delta_i|$, the α_i and β_i referring to rows and columns in both cases, the γ_i as treatments in the first Latin square, and δ_i as treatments on the second Latin square, such that when the two Latin squares are superimposed each γ_i will occur with each δ_i exactly once. Two Latin squares which have this property are said to be *orthogonal*. A set of r - 1 mutually orthogonal Latin squares

is said to form a complete set of mutually orthogonal Latin squares and when superimposed would form an (r+1)-way classification Graeco-Latin square. Complete sets of orthogonal Latin squares exist when r is a prime integer and also for certain other values, e. g. r = 4,8,9. The sum of squares S in the likelihood function is

$$S = \sum_{i,j}(y_{ij}-m-R_i-C_j-T_t-U_u)^2,$$

and may be written as

$$S = S_E + S_R + S_C + S_T + S_U + r^2(\bar{y}-m)^2,$$

where

$$S_E = \sum_{i,j}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}-\bar{y}_{(t)}-\bar{y}_{[u]}+3\bar{y})^2,$$
$$S_R = \sum_{i,j}(\bar{y}_{i.}-\bar{y}-R_i)^2,$$
$$S_C = \sum_{i,j}(\bar{y}_{.j}-\bar{y}-C_j)^2,$$
$$S_T = \sum_{i,j}(\bar{y}_{(t)}-\bar{y}-T_t)^2,$$
$$S_U = \sum_{i,j}(y_{[u]}-\bar{y}-U_u)^2,$$

where $\bar{y}_{i.}$, $\bar{y}_{.j}$, $\bar{y}$, $\bar{y}_{(t)}$ are as defined in §9.4 and $\bar{y}_{[u]}$ is the average of all y_{ij} having mean values involving U_u. Let S_R^o be the value of S_R when the $R_i = 0$ with similar meanings for S_C^o, S_T^o and S_U^o.

As before, we may define hypotheses $H[(R_i)=0]$, $H[(C_j)=0]$, $H[(T_t)=0]$, and $H[(U_u)=0]$ all with the same Ω parameter space given by

$$\Omega: \begin{cases} -\infty < m, R_i, C_j, T_t, U_u < +\infty, \sigma^2 > 0, \\ \sum R_i = \sum C_j = \sum T_t = \sum U_u = 0, \end{cases}$$

but with ω parameter spaces obtained by setting each $R_i = 0$, each $C_j = 0$, each $T_t = 0$, and each $U_u = 0$, respectively. The F ratios for these four hypotheses may be written down by the reader in terms of S_E, S_R^o, S_C^o, S_T^o, and S_U^o.

The analysis of variance table for the four-way Graeco-Latin square turns out to be as follows:

Variation Due to	Sum of Squares	Degrees of Freedom
The α_i	$S_R^0 = r\sum_i (\bar{y}_{i\cdot} - \bar{y})^2$	$r - 1$
The β_j	$S_C^0 = r\sum_j (\bar{y}_{\cdot j} - \bar{y})^2$	$r - 1$
The γ_t	$S_T^0 = r\sum_{(t)} (\bar{y}_{(t)} - \bar{y})^2$	$r - 1$
The δ_u	$S_U^0 = r\sum_{[u]} (\bar{y}_{[u]} - \bar{y})^2$	$r - 1$
Error	$S_E = \sum_{i,j} (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} - \bar{y}_{(t)} - \bar{y}_{[u]} + 3\bar{y})^2$	$(r-1)(r-3)$
Total	$S = \sum_{i,j} (y_{ij} - \bar{y})^2$	$r^2 - 1$

where $\bar{y}$, $\bar{y}_{i\cdot}$, $\bar{y}_{\cdot j}$, $\bar{y}_{(t)}$ have the same meanings as for the Latin Square and $y_{[u]}$ is the mean of all y_{ij} having attribute δ_u.

The properties of the constituents of this table are very similar to those of the constituents of the table pertaining to the ordinary Latin Square and therefore we shall not write them down. The reader may verify that S_E is the minimum of $\sum_{i,j=1}^{r} (y_{ij} - m - R_i - C_j - T_t - U_u)^2$, subject to the restrictions $\sum_1^r R_i = \sum_1^r C_j = \sum_1^r T_t = \sum_1^r U_u = 0$, and is obtainable from formula (k), §8.5.

Extensions to higher order Graeco-Latin squares and complete sets of Latin squares are straightforward.

9.6 Analysis of Variance in Incomplete Layouts

The results which have been presented in §§9.2-9.5 depend on complete or balanced layouts in the sense that there is exactly one random variable corresponding to each cell of the layout, or in the sense of orthogonality exemplified by Latin Squares, Graeco-Latin squares, and complete sets of Latin squares. Because of this element of balance the sums of squares arising in connection with the various hypotheses are relatively simple. The problem to be considered here is that of deriving sums of squares appropriate to tests of hypotheses in case there are arbitrary numbers of random variables associated with the various cells.

First let us consider the case of a two-way layout. Let $y_1, y_2, \ldots, y_n$ be the random variables of the sample such that each y belongs to one row and one column in an r by s layout. If a y, say y_α, belongs to the i-th row and j-th column, we assume it to be distributed according to $N(m+R_i+C_j, \sigma^2)$ where $\sum_i R_i = \sum_j C_j = 0$. We may rewrite this distribution as $N(m + \sum_i R_i x_{i1\alpha} + \sum_j C_j x_{2j\alpha}, \sigma^2)$ where for a given α the $x_{i1\alpha}$ $(i = 1,2,\ldots,$

are all zero except for the value of i corresponding to the row within which y_α occurs, a and similarly the $x_{2j\alpha}$ ($j = 1,2,\ldots,s$) are all zero except for the value of j corresponding to the column within which y_α occurs.

The likelihood function for the sample $y_1,\ldots,y_n$ is

$$\text{(a)} \qquad \prod_{\alpha=1}^{n} N(m+\sum_i R_i x_{1i\alpha}+\sum_j C_j x_{2j\alpha},\sigma^2),$$

and the sum of squares in the exponent of this likelihood function is

$$\text{(b)} \qquad S = \sum_{\alpha=1}^{n}(y_\alpha - m - \sum_i R_i x_{1i\alpha} - \sum_j C_j x_{2j\alpha})^2 .$$

Now suppose we consider the hypothesis that the C_j are all 0. This hypothesis $H[(C_j)=0]$ may be specified as follows:

$$\text{(c)} \qquad \Omega: \begin{cases} -\infty < m,\ R_i,\ C_j < \infty,\ \sigma^2 > 0, & \text{(all i and j)} \\ \sum_i R_i = \sum_j C_j = 0 , \end{cases}$$

$$\omega: \quad [\ \text{Subspace in } \Omega \text{ obtained by setting each } C_j = 0.$$

Maximizing the likelihood function for variations of the parameters in Ω we find from §8.5 that the values of m, the R_i and C_j which minimize S are given by the linear equations

$$\text{(d)} \qquad \begin{aligned} -\sum_{\cdot\cdot} y_\alpha + nm + \sum_i n_{i\cdot} R_i + \sum_j n_{\cdot j} C_j &= 0, \\ -\sum_{i\cdot} y_\alpha + n_{i\cdot} m + n_{i\cdot} R_i + \sum_j n_{ij} C_j + \lambda_1 &= 0, \qquad i = 1,2,\ldots,r, \\ -\sum_{\cdot j} y_\alpha + n_{\cdot j} m + \sum_i n_{ij} R_i + n_{\cdot j} C_j + \lambda_2 &= 0, \qquad j = 1,2,\ldots,s, \\ \sum_i R_i &= 0, \\ \sum_j C_j &= 0, \end{aligned}$$

where $\sum_{\cdot\cdot} y_\alpha$ denotes summation of all y_α, $\sum_{i\cdot} y_\alpha$ denotes summation of all y_α in the i-th row, $\sum_{\cdot j} y_\alpha$ denotes summation of all y_α in the j-th column, n_{ij} is the number of y_α falling in the cell at the intersection of the i-th row and j-th column, $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$. It follows from §8.5 that the minimum of S for variations of the m, R_i and C_j in Ω is given by

$$\min_\Omega(S) = \frac{\Delta}{\Delta_{oo}} ,$$

where

(e) $$\Delta = \begin{vmatrix} \sum_{..} y_\alpha^2 & \sum_{..} y_\alpha & \sum_{1.} y_\alpha \cdots \sum_{r.} y_\alpha & \sum_{.1} y_\alpha \cdots \sum_{.s} y_\alpha & 0 & 0 \\ \sum_{..} y_\alpha & n & n_{1.} \cdots n_{r.} & n_{.1} \cdots n_{.s} & 0 & 0 \\ \sum_{1.} y_\alpha & n_{1.} & n_{1.} \cdots 0 & n_{11} \cdots n_{1s} & 1 & 0 \\ \vdots & \vdots & \vdots \quad \vdots & \vdots \quad \vdots & \vdots & \vdots \\ \sum_{r.} y_\alpha & n_{r.} & 0 \cdots n_{r.} & n_{r1} \cdots n_{rs} & 1 & 0 \\ \sum_{.1} y_\alpha & n_{.1} & n_{11} \cdots n_{r1} & n_{.1} \cdots 0 & 0 & 1 \\ \vdots & \vdots & \vdots \quad \vdots & \vdots \quad \vdots & \vdots & \vdots \\ \sum_{.s} y_\alpha & n_{.s} & n_{1s} \cdots n_{rs} & 0 \cdots n_{.s} & 0 & 1 \\ 0 & 0 & 1 \cdots 1 & 0 \cdots 0 & 0 & 0 \\ 0 & 0 & 0 \cdots 0 & 1 \cdots 1 & 0 & 0 \end{vmatrix}$$

and Δ_{00} is the minor of $\sum_{..} y_\alpha^2$ in Δ. Hence

$$\hat{\sigma}_\Omega^2 = \frac{1}{n} \frac{\Delta}{\Delta_{00}} .$$

Maximizing the likelihood function for variations of the parameters over ω we find that the maximizing values of m and the R_i are given by the $r + 2$ equations resulting by setting λ_2 and all C_j equal to zero in (d) and deleting the last equation. Similarly,

$$\min_\omega(S) = \frac{\Delta'}{\Delta'_{00}},$$

where Δ' is obtained by deleting the last $s + 2$ rows and columns from Δ with exception of the next to the last row and column. Δ'_{00} is the minor of $\sum_{..} y_\alpha^2$ in Δ'.
Hence

$$\hat{\sigma}_\omega^2 = \frac{1}{n} \frac{\Delta'}{\Delta'_{00}} .$$

It follows from Theorem (A), §8.3, that

$$q_1 = \frac{n\hat{\sigma}_\Omega^2}{\sigma^2} = \frac{\Delta}{\sigma^2 \Delta_{00}},$$

and

$$q_2 = \frac{n(\hat{\sigma}_\omega^2 - \hat{\sigma}_\Omega^2)}{\sigma^2} = \frac{1}{\sigma^2} \left(\frac{\Delta'}{\Delta'_{00}} - \frac{\Delta}{\Delta_{00}} \right),$$

are distributed independently according to χ^2-laws with n - r - s + 1 and s - 1 degrees of freedom respectively when $H[(C_j)=0]$ is true. The F ratio is therefore

$$(f) \qquad \frac{(n-r-s+1)\left(\frac{\Delta'}{\Delta_{oo}} - \frac{\Delta}{\Delta_{oo}}\right)}{(s-1)\left(\frac{\Delta}{\Delta_{oo}}\right)},$$

which has the distribution $h_{(s-1),(n-r-s+1)}(F)dF$ when $H[(C_j)=0]$ is true. The reader may verify that if $n_{ij} = 1$ (all i and j) then n = rs and we have the complete two-way layout discussed in §9.2, and in this case the F ratio reduces to that given in (b), §9.2.

The extension of the foregoing treatment to higher order layouts is straightforward and will not be considered in detail. It is perhaps sufficient to note that in the case of higher-order layouts we would have several sets, say q, classifications, the u-th classification consisting of p_u mutually exclusive categories, such that each y_α in the sample would belong to exactly one category in each classification. If we denote the mean effect (on y_α) of the v-th category of the u-th classification by I_{uv} where $\sum_{v=1}^{p_u} I_{uv} = 0$, u = 1,2,...,q (or more generally several linear restrictions may be applied to I_{uv} for each u) then the mean value of y_α may be expressed as $m + \sum_{u=1}^{q} \sum_{v=1}^{p_u} I_{uv} x_{uv\alpha}$ where for each value of u, $x_{uv\alpha}$ (α = 1,2,...,n) is unity for only one value of v and zero otherwise; the value of v for which $x_{uv\alpha}$ is unity being that corresponding to the category (of the u-th classification) within which y_α falls. The problem of testing the hypothesis that I_{uv} for the u-th classification (u-th classification <u>effects</u>) are all zero amounts to setting up a determinant corresponding to Δ in (e) based on q classifications instead of 2, and performing operations similar to those performed on Δ to obtain Δ_{oo}, Δ', and Δ'_{oo}. The reader will find it instructive to work through the details of setting up Δ, q_1, q_2, and F for the case of a three-way layout when the hypothesis to be tested is that the main effects due to one of the classifications are zero. He will also find it profitable to treat the ordinary Latin square as a three-way layout by this method and show that the F obtained for testing the hypothesis of no treatment effects is identical with that given by (d) in §9.4. The generality of this procedure should be carefully noted by the reader because not only can all of the results previously discussed in this chapter be obtained by this procedure, but tests for the existence of interaction between two or more classifications in incomplete or unbalanced layouts may be deduced by applying the procedure.

9.7 Analysis of Covariance

Throughout all of the discussion in §§9.2-9.6 we have assumed the mean value of the random variable in each case to consist of the sum of a general constant (which is the same for all random variables) and constants referring to rows, columns, treatments, in-

teraction, etc. It frequently happens that there are practical situations which suggest that the mean value of the random variable should include linear functions of one or more fixed variates (see §8.2) in addition to the sum of constants of the type mentioned above. For example, if y_{ij} refers to yield of wheat in a plot in the i-th row and j-th column of a two-way layout, not only should the mean value of y_{ij} include a general constant and row and column effects, but also linear effect of number of plants on this plot, say x_{ij}. The mean value would then be of the form $m + ax_{ij} + R_i + C_j$, with the usual conditions on the R_i and C_j. The object of this section is to examine what modifications of §§9.2-9.6 should be made in order to take one or more fixed variates into account in the mean value of the random variables involved.

Let us return to the two-way layout discussed in §9.2 and assume that the mean value of y_{ij} depends linearly not only on m, R_i and C_j but also on a fixed variable x_{ij}. In other words, assume that the y_{ij} are random variables independently distributed according to $N(m+ax_{ij}+R_i+C_j, \sigma^2)$, where $\sum_i R_i = \sum_j C_j = 0$. The question arises as to what forms the F-ratios take for testing the hypothesis that the C_j are all zero or the hypothesis that the R_i are all zero, when the Ω parameter space is the $(r + s + 1)$-dimensional space for which $-\infty < a, m, R_i, C_j < +\infty$, $\sigma^2 > 0$. The probability element of the y_{ij} is exactly that given in (a), §9.2, with y_{ij} replaced by $y_{ij} - ax_{ij}$. Making this substitution in (b), §9.2, we see that the sum of squares in the exponent of this probability element (for any point in Ω) may be broken down into the following components:

$$\text{(a)} \qquad S_\Omega = \sum_{i,j}(Y_{ij}-aX_{ij})^2 + \sum_{i,j}(\bar{Y}_{i\cdot}-a\bar{X}_{i\cdot}-R_i)^2 + \sum_{i,j}(\bar{Y}_{\cdot j}-a\bar{X}_{\cdot j}-C_j)^2 + rs(\bar{y}-a\bar{x}-m)^2,$$

where $Y_{ij} = y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}$, with similar meaning for X_{ij}, and $Y_{i\cdot} = \bar{y}_{i\cdot}-\bar{y}$, with similar meaning for $\bar{Y}_{\cdot j}$, $\bar{X}_{i\cdot}$, $\bar{X}_{\cdot j}$. The first sum of squares on the right in (a) may be written as

$$\text{(b)} \qquad \sum_{i,j}(Y_{ij}-aX_{ij})^2 = \sum_{i,j}(Y_{ij}-\hat{a}\,X_{ij})^2 + (\hat{a}-a)^2\sum_{i,j}X_{ij}^2,$$

where $\hat{a}_\Omega = \dfrac{\sum_{i,j}X_{ij}Y_{ij}}{\sum_{i,j}X_{ij}^2}$.

Making the substitution (b) in (a) we obtain 5 sums of squares which when divided by σ^2 are (by Cochran's theorem) distributed independently according to χ^2-laws with $(r-1)(s-1)$, -1, 1, $r-1$, $s-1$, 1 degrees of freedom, respectively.

Now suppose we wish to test the hypothesis $H_1[(C_j)=0]$ which is specified as follows:

(c)
$$\Omega: \begin{cases} -\infty < a,\ m,\ R_i,\ C_j < \infty,\ \sigma^2 > 0, \text{ (all } i,j) \\ \sum_i R_i = \sum_j C_j = 0. \end{cases}$$

ω: { The subspace in Ω obtained by setting each $C_j = 0$.

Maximizing the likelihood function for variations of the parameters in Ω, which is equivalent to minimizing S as far as variations of a, m, R_i, C_j are concerned, we obtain

(d) $$\hat{R}_{i,\Omega} = \bar{Y}_{i\cdot} - \hat{a}_\Omega \bar{X}_{i\cdot},\quad \hat{C}_{j,\Omega} = \bar{Y}_{\cdot j} - \hat{a}_\Omega \bar{X}_{\cdot j},\quad \hat{m}_\Omega = \bar{y} - \hat{a}_\Omega \bar{x},\quad \sigma^2_\Omega = \frac{1}{rs}\sum_{i,j}(Y_{ij} - \hat{a}_\Omega X_{ij})^2.$$

The sum of squares in the exponent of the probability element for any point in ω (i. e., all $C_j = 0$) may be expressed in terms of the following components:

(e) $$S_\omega = \sum_{i,j}[(Y_{ij}+\bar{Y}_{\cdot j}) - a(X_{ij}+\bar{X}_{\cdot j})]^2 + \sum_{i,j}(\bar{Y}_{i\cdot} - a\bar{X}_{i\cdot} - R_i)^2 + rs(\bar{y} - a\bar{x} - m)^2.$$

Maximizing the likelihood function for variations of the parameters in ω amounts to minimizing S_ω as far as m, a, R_i are concerned. We find

$$\hat{a}_\omega = \frac{\sum_{ij}(Y_{ij}+\bar{Y}_{\cdot j})(X_{ij}+\bar{X}_{\cdot j})}{\sum_{ij}(X_{ij}+\bar{X}_{\cdot j})^2},$$

(f) $$\hat{R}_{i,\omega} = \bar{Y}_{i\cdot} - \hat{a}_\omega \bar{X}_{i\cdot},\qquad \hat{m}_\omega = \bar{Y} - \hat{a}_\omega \bar{X},$$

$$\hat{\sigma}^2_\omega = \frac{1}{rs}\sum_{i,j}[(Y_{ij}+\bar{Y}_{\cdot j}) - \hat{a}_\omega(X_{ij}+\bar{X}_{\cdot j})]^2$$

$$= \frac{1}{rs}\Big\{\sum_{i,j}(Y_{ij} - \hat{a}_\omega X_{ij})^2 + \sum_{ij}(\bar{Y}_{\cdot j} - \hat{a}_\omega \bar{X}_{\cdot j})^2\Big\}.$$

By Theorem (A), §8.3, it follows that

$$q_1 = \frac{rs\hat{\sigma}^2_\Omega}{\sigma^2},\qquad q_2 = \frac{rs(\hat{\sigma}^2_\omega - \hat{\sigma}^2_\Omega)}{\sigma^2}$$

are independently distributed according to χ^2-laws with $(r-1)(s-1)-1$ and $s-1$ degrees of freedom, respectively, when $H_1[(C_j)=0]$ is true. Hence the F-ratio for this hypothesis is

$$\frac{[(r-1)(s-1)-1](\hat{\sigma}^2_\omega - \hat{\sigma}^2_\Omega)}{(s-1)\hat{\sigma}^2_\Omega},$$

which has the distribution $h_{s-1,(r-1)(s-1)-1}(F)dF$ when the hypothesis is true.

It should be noted that $rs\hat{\sigma}^2_\Omega$ and $rs\hat{\sigma}^2_\omega$ can each be expressed in terms of determinants (see (g), §8.2) as follows:

$$rs\ \hat{\sigma}_{\Omega}^2 = \begin{vmatrix} \sum_{i,j} Y_{ij}^2 & \sum_{i,j} X_{ij}Y_{ij} \\ \sum_{i,j} X_{ij}Y_{ij} & \sum_{i,j} X_{ij}^2 \end{vmatrix} \Big/ \sum_{i,j} X_{ij}^2,$$

$$rs\ \hat{\sigma}_{\omega}^2 = \begin{vmatrix} \sum_{i,j} (Y_{ij}+\bar{Y}_{\cdot j})^2 & \sum_{i,j} (X_{ij}+\bar{X}_{\cdot j})(Y_{ij}+\bar{Y}_{\cdot j}) \\ \sum_{i,j} (X_{ij}+\bar{X}_{\cdot j})(Y_{ij}+\bar{Y}_{\cdot j}) & \sum_{i,j} (X_{ij}+\bar{X}_{\cdot j})^2 \end{vmatrix} \Big/ \sum_{i,j} (X_{ij}+\bar{X}_{\cdot j})^2$$

$$= \begin{vmatrix} \sum_{i,j} Y_{ij}^2 + \sum_{i,j} \bar{Y}_{\cdot j}^2 & \sum_{i,j} X_{ij}Y_{ij} + \sum_{i,j} \bar{X}_{\cdot j}\bar{Y}_{\cdot j} \\ \sum_{i,j} X_{ij}Y_{ij} + \sum_{i,j} \bar{X}_{\cdot j}\bar{Y}_{\cdot j} & \sum_{i,j} X_{ij}^2 + \sum_{i,j} \bar{X}_{\cdot j}^2 \end{vmatrix} \Big/ \left(\sum_{i,j} X_{ij}^2 + \sum_{i,j} \bar{X}_{\cdot j}^2\right).$$

In a similar manner we may define hypothesis $H_1[(R_i)=0]$ by replacing $C_j = 0$ by $R_i = 0$ in the specification of ω. We find that $\hat{\sigma}_{\Omega}^2$ remains the same but $\hat{\sigma}_{\omega}^2$ for this hypothesis is identical with that for $H_1[(C_j)=0]$ after replacing $\bar{Y}_{\cdot j}$ and $\bar{X}_{\cdot j}$ by $\bar{Y}_{i\cdot}$ and $\bar{X}_{i\cdot}$, respectively. The F-ratio for $H_1[(R_i)=0]$ is distributed according to $h_{r-1,\,(r-1)(s-1)-1}(F)dF$.

The constituents which are used in making up the F-ratios for testing the two hypotheses considered above may be set forth in the following analysis of covariance table:

Variation Due To	Sums of Squares		Cross Products	Degrees of Freedom
	y	x	(x) (y)	
Rows	$\sum_{i,j} \bar{Y}_{i\cdot}^2$	$\sum_{i,j} \bar{X}_{i\cdot}^2$	$\sum_{i,j} \bar{X}_{i\cdot}\bar{Y}_{i\cdot}$	r - 1
Columns	$\sum_{i,j} \bar{Y}_{\cdot j}^2$	$\sum_{i,j} \bar{X}_{\cdot j}^2$	$\sum_{i,j} \bar{X}_{\cdot j}\bar{Y}_{\cdot j}$	s - 1
Error	$\sum_{i,j} Y_{ij}^2$	$\sum_{i,j} X_{ij}^2$	$\sum_{i,j} X_{ij}Y_{ij}$	(r-1)(s-1)
Total	$\sum_{i,j} (y_{ij}-\bar{y})^2$	$\sum_{i,j} (x_{ij}-\bar{x})^2$	$\sum_{i,j} (y_{ij}-\bar{y})(x_{ij}-\bar{x})$	rs - 1

The results obtained for the case of one fixed variate may be extended in a rather straight forward manner to the case of k fixed variates where $k < (r-1)(s-1)$. Thus, if k fixed variates x_{pij}, $p = 1,2,\ldots,k$, are taken into account linearly in our two-way layout, we would begin by replacing y_{ij} in the probability element in (a), §9.2, by $(y_{ij} - \sum_{p=1}^{k} a_p x_{pij})$ and follow a procedure similar to that for the case of one fixed variate. Thus, in place of aX_{ij}, $a\bar{X}_{i\cdot}$, $a\bar{X}_{\cdot j}$, $a\bar{X}$ in (a) we would have $\sum_p a_p X_{pij}$, $\sum_p a_p \bar{X}_{pi\cdot}$, $\sum_p a_p \bar{X}_{p\cdot j}$, $\sum_p a_p \bar{X}_p$, respectively, where the meanings of X_{pij}, $\bar{X}_{pi\cdot}$, $\bar{X}_{p\cdot j}$, $\bar{X}_p$ are obvious. The reader will find it instructive to carry out the details in arriving at F-ratios for

testing hypotheses $H_k[(C_j)=0]$ and $H_k[(R_i)=0]$ which are k-fixed-variate analogues of $H_1[(C_j)=0]$ and $H_1[(R_i=0]$, respectively.

The procedure which we have outlined for introducing fixed variates linearly into the mean value of the random variables in a two-way layout extends in a straight forward manner to three-way layouts, Latin squares, Graeco-Latin squares, and to incomplete or non-orthogonal layouts of the type discussed in §9.6. We shall have to leave the matter of carrying out details as exercises for the reader. Because of the generality of §9.6 it is perhaps worth while to remark, without going through the details of proof, that if one fixed variate is introduced linearly into the mean value of y_α, which would amount to replacing m by $m + ax_\alpha$ in (a), §9.6, the effect on the determinant Δ as defined in (e), §9.6, would be to insert another row and column into Δ as second row and second column, the $r + s + 5$ elements of this row and column being

$$\sum_{\cdot\cdot} x_\alpha y_\alpha, \sum_{\cdot\cdot} x_\alpha^2, \sum_{\cdot\cdot} x_\alpha, \sum_{1\cdot} x_\alpha, \ldots\ldots, \sum_{r\cdot} x_\alpha, \sum_{\cdot 1} x_\alpha, \sum_{\cdot 2} x_\alpha, \ldots, \sum_{\cdot s} x_\alpha, 0, 0$$

reading left to right in the row, and reading top to bottom in the column. This augmented determinant has its own Δ_{00}, Δ', Δ'_{00} (see §9.6) which are obtained by operations analogous to those used in obtaining Δ_{00}, Δ', Δ'_{00} from Δ in §9.6. The extension of our procedure to the problem of linearly taking into account k fixed variates in the mean value of y_α in §9.6 is straightforward and will be left to the reader.

CHAPTER X

ON COMBINATORIAL STATISTICAL THEORY

Many problems in distribution or sampling theory in statistics reduce to combinatorial considerations. For example, the derivation of the binomial distribution (§3.11) depends on the determination of the number of distinct orders in which x p's and n-x q's can be multiplied together, and similarly the derivation of the multinomial distribution (§3.12) depends on the enumeration of the number of distinct orders in which n_1 p_1's, n_2 p_2's,...,n_k p_k's can be multiplied together where $\sum_{i=1}^{k} p_k = 1$, $\sum_{i=1}^{k} n_1 = n$. A majority of the combinatorial problems of the drawing-balls-from-urns variety involve direct applications of permutation and combination formulas, which in turn are often simply expressible in terms of binomial and multinomial coefficients. The theory of sampling from a finite population (§4.3) is based on the use of binomial and multinomial coefficients and their use as weights in various averaging operations. The sampling theory of order statistics (§4.5) is a direct application of the multinomial distribution law to probability functions of continuous random variables.

The object of the present chapter is to discuss some of the more complicated distribution problems in combinatorial statistical theory which are of particular interest in applied mathematical statistics. More specifically, we shall present some results on the theory of runs, the theory of matching and its application to testing independence in contingency tables, Pearson's original χ^2-problem, and inspection sampling.

10.1 On the Theory of Runs

Suppose we have an arbitrary sequence of n elements, each element being one of several mutually exclusive kinds. Each sequence of elements of one kind is called a run. The simplest case is that in which there are two kinds of objects. We shall consider this case in detail, and also present briefly some results for the case of several kinds of elements.

10.11 Case of Two Kinds of Elements

Suppose we have n_1 a's and n_2 b's ($n_1+n_2=n$). Let r_{1j} denote the number of runs of a's of length j and r_{2j} denote the number of runs of b's of length j. For example, if

the arrangement is

aaabbaabaabbab,

then $r_{11} = 1$, $r_{12} = 2$, $r_{13} = 1$, $r_{21} = 2$, $r_{22} = 2$, and the other r's are zero. It should be observed that $\sum_j j r_{1j} = n_1$, the number of a's, and also $\sum_j j r_{2j} = n_2$. Let $r_1 = \sum_j r_{1j}$ and $r_2 = \sum_j r_{2j}$ denote the total number of runs of a's and b's, respectively. For a given set of numbers $r_{11}, r_{12}, r_{13}, \ldots$ there are $\frac{r_1!}{r_{11}!r_{12}!\cdots\cdots r_{1n_1}!}$ ways of arranging the r_1 runs of a's. And for a specified set, r_{2j}, there are $\frac{r_2!}{r_{21}!r_{22}!\cdots\cdots r_{2n_2}!}$ ways of arranging r_2 runs of b's. It is clear that r_1 cannot differ from r_2 by more than unity, for if it did two runs of one kind of element would have to be adjacent, but this is contrary to the definition of runs. If $r_1 = r_2$, a given arrangement of runs of a's can be fitted into a given arrangement of runs of b's in two ways, either with a run of a's first or with a run of b's first. We define the function $F(r_1, r_2)$ to be the number of ways of arranging r_1 objects of one kind and r_2 objects of another so that no two adjacent objects are of the same kind. Clearly,

(a)
$$\begin{aligned} F(r_1,r_2) &= 0 \text{ if } |r_1-r_2| > 1 \\ &= 1 \text{ if } |r_1-r_2| = 1 \\ &= 2 \text{ if } r_1 = r_2. \end{aligned}$$

The total number of ways of getting the set r_{ij} $(i = 1,2;\ j = 1,2,\ldots,n_i)$ is

$$N(r_{ij}) = \frac{r_1!}{r_{11}!r_{12}!\cdots r_{1n_1}!}\ \frac{r_2!}{r_{21}!r_{22}!\cdots r_{2n_2}!} \cdot F(r_1,r_2).$$

Since there are $\frac{n!}{n_1!n_2!}$ possible arrangements of a's and b's, the joint distribution function of the given set r_{ij} (all possible arrangements given equal weight) is

(b)
$$p(r_{ij}) = \frac{r_1!}{r_{11}!\cdots r_{1n_1}!} \cdot \frac{r_2!}{r_{21}!\cdots r_{2n_2}!} \cdot F(r_1,r_2) \Big/ \frac{n!}{n_1!n_2!}.$$

Now let us determine the joint distribution of the r_{1j}. To do this we sum $p(r_{ij})$ with respect to the r_{2j}. We wish to sum $\frac{r_2!}{r_{21}!\cdots r_{2n_2}!}$ over all partitions of n_2, i.e., for all r_{2j} such that $\sum_{j=1}^{n_2} j r_{2j} = n_2$ and $\sum_j r_{2j} = r_2$. In order to do this, consider

$$(x+x^2+x^3+\ldots)^{r_2} = \frac{x^{r_2}}{(1-x)^{r_2}}$$

$$= x^{r_2} \sum_{t=0}^{\infty} \frac{(r_2-1+t)!}{(r_2-1)!t!} x^t.$$

It is evident that the coefficient of x^{n_2} in the initial expression is the sum $\sum \frac{r_2!}{r_{21}!\ldots r_{2n_2}!}$ that we desire. The coefficient of x^{n_2} in the final expression is the coefficient of the term for which $r_2+t = n_2$, i. e., $t = n_2-r_2$. Therefore the desired sum is $\frac{(r_2-1+n_2-r_2)!}{(r_2-1)!(n_2-r_2)!} = \frac{(n_2-1)!}{(r_2-1)!(n_2-r_2)!}$. Hence, the joint distribution function of the r_{1j} and r_2 is

$$(c) \qquad p(r_{1j},r_2) = \frac{r_1!}{r_{11}!\ldots r_{1n_1}!} \cdot \frac{(n_2-1)!}{(r_2-1)!(n_2-r_2)!} \cdot F(r_1,r_2) \Big/ \frac{n!}{n_1!n_2!} .$$

Now we sum out r_2. By (a) we get

$$\sum_{r_2=1}^{n_2} \frac{(n_2-1)!}{(r_2-1)!(n_2-r_2)!} F(r_1,r_2) = \frac{(n_2-1)!}{(r_1-2)!(n_2-r_1+1)!} \cdot 1 + \frac{(n_2-1)!}{(r_1-1)!(n_2-r_1)!} \cdot 2$$

$$+ \frac{(n_2-1)!}{r_1!(n_2-r_1-1)!} \cdot 1 = \frac{(n_2+1)!}{r_1!(n_2-r_1+1)!} .$$

This gives us the joint distribution function of the r_{1j}

$$(d) \qquad p(r_{1j}) = \frac{r_1!}{r_{11}!r_{12}!\ldots r_{1n_1}!} \cdot \frac{(n_2+1)!}{r_1!(n_2-r_1+1)!} \Big/ \frac{n!}{n_1!n_2!} ,$$

with a similar expression holding for the joint distribution of the r_{2j}.

Another important distribution is the joint distribution of r_1 and r_2. We get this by summing out the r_{1j} in (c), just as we summed (b) with respect to the r_{2j} to obtain (c). The result is

$$(e) \qquad p(r_1,r_2) = \frac{(n_1-1)!}{(r_1-1)!(n_1-r_1)!} \cdot \frac{(n_2-1)!}{(r_2-1)!(n_2-r_2)!} \cdot F(r_1,r_2) \Big/ \frac{n!}{n_1!n_2!} .$$

Finally, we find the distribution function of r_1 by summing (e) with respect to r_2, obtaining

$$(f) \qquad p(r_1) = \frac{(n_1-1)!}{(r_1-1)!(n_1-r_1)!} \frac{(n_2+1)!}{r_1!(n_2+1-r_1)!} \Big/ \frac{n!}{n_1!n_2!} .$$

The distribution of the total number of runs of a's and b's is of considerable interest in applications of run theory. It is used as a test for randomness of the arrangement of a's and b's; the smaller the total number of runs the more untenable the hypothesis of randomness. Let $u = r_1 + r_2$, the total number of runs. To find the distribution of u we must sum (e) over all points in the r_1, r_2 plane for which $u = r_1 + r_2$. We have two cases, (1) $u = 2k$ (even) and (2) $u = 2k-1$ (odd). To find the probability that

$u = 2k$, we note there is only one point in the r_1, r_2 plane for which $u = r_1 + r_2 = 2k$ where $F(r_1, r_2) \neq 0$, and that point is (k,k). When $u = r_1 + r_2 = 2k - 1$ there are two points at which $F(r_1, r_2) \neq 0$, namely $(k, k-1)$ and $(k-1, k)$. Hence from (e) we have at once (using the notation ${}_mC_n = \binom{m}{n}$):

$$\text{(g)} \qquad \Pr(u=2k) = 2\binom{n_1-1}{k-1}\cdot\binom{n_2-1}{k-1}\Big/\binom{n_1+n_2}{n_1},$$

$$\Pr(u=2k-1) = \frac{\binom{n_1-1}{k-1}\cdot\binom{n_2-1}{k-2} + \binom{n_1-1}{k-2}\cdot\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}.$$

This distribution was derived by Stevens* and also by Wald and Wolfowitz** and the function $\Pr(u \leq u') = \sum_{u=2}^{u'} p(u)$ has been tabulated by Swed and Eisenhart*** for $n_1 \leq n_2$ (n_1=m, n_2=n in their notation) from the case $n_1 = 2$, $n_2 = 20$ to $n_1 = 19$, $n_2 = 20$ for various values of u'.

Another probability function of considerable interest in the application of the theory of runs is the probability of getting at least one run of a's of length s or greater or in other words the probability that at least one of the variables r_{1s}, r_{1s+1}, $r_{1s+2}, \ldots$, in the distribution (d) is ≥ 1. Mosteller**** has solved this problem for the case $n_1 = n_2 = n$. To obtain this probability we put $n_1 = n_2 = n$ in (d), thus obtaining

$$\text{(h)} \qquad p(r_{1j}) = \frac{r_1!}{r_{11}!r_{12}!\ldots r_{1n}!}\,\frac{(n+1)!}{r_1!(n-r_1+1)!}\Big/\frac{(2n)!}{n!^2},$$

and sum over all terms such that at least one of the variables $r_{1s}, r_{1s+1}, \ldots \geq 1$. We can accomplish the same thing by summing over all terms such that all of these variables are zero, and subtracting the result from unity. To do this we must sum the multinomial coefficient in (h) over all values of $r_{11}, \ldots, r_{1n}$ such that $r_{1s} = r_{1s+1} = \ldots = r_{1n} = 0$,

---

*W. L. Stevens, "Distribution of Groups in a Sequence of Alternatives", Annals of Eugenics, Vol. IX (1939).

**A. Wald and J. Wolfowitz, "On a Test of Whether Two Samples are from the Same Population", Annals of Math. Stat., Vol. XI (1940).

***Frieda S. Swed and C. Eisenhart, "Tables for Testing Randomness of Grouping in a Sequence of Alternatives", Annals of Math. Stat., Vol. XIV (1943).

****Frederick Mosteller, "Note on an Application of Runs to Quality Control Charts", Annals of Math. Stat., Vol. XII (1941).

$\sum_1^n jr_{1j} = n$, $\sum_1^n r_{1j} = r_1$, and then sum with respect to r_1. It will be noted that the sum of the multinomial coefficients under these conditions is given by the coefficient of x^n in the formal expansion of

$$(x+x^2+\ldots+x^{s-1})^{r_1} = x^{r_1}(1-x^{s-1})^{r_1}\sum_{t=0}^{\infty}\binom{r_1-1+t}{r_1-1}x^t,$$

which is

$$\sum_{j=0}^{r_1}(-1)^j\binom{r_1}{j}\binom{n-j(s-1)-1}{r_1-1}.$$

The desired probability of at least one run of length s or greater is therefore

$$\text{(i)}\qquad \Pr(\text{at least one of } r_{1j}>1,\ j\geq s) = 1 - \frac{\sum_{r_1}\sum_{j=0}^{r_1}(-1)^j\binom{r_1}{j}\binom{n-1-j(s-1)}{r_1-1}\binom{n+1}{r_1}}{\binom{2n}{n}},$$

the summation on r_1 extending from ~~1~~ the largest integer $\leq \frac{n}{s-1}$ ~~+1~~. Applying similar methods to each of the multinomial coefficients in (b), Mosteller has shown that the probability of getting at least one run of a's or b's of length s or greater is

$$\text{(j)}\qquad \Pr(\text{at least one of } r_{1j} \text{ or } r_{2j} > 1,\ j \geq s) = 1 - A/\binom{2n}{n},$$

where

$$A = \sum_{r_1}\left\{\sum_{r_2=r_1-1}^{r_1+1}F(r_1,r_2)\left[\prod_{i=1}^{2}\sum_{j=1}^{r_i}(-1)^j\binom{r_i}{j}\binom{n-1-j(s-1)}{r_i-1}\right]\right\},$$

the r_1 summation being similar to that in (i). Mosteller has tabulated the smallest value of s for which each of the probabilities (i) and (j) is $\geq .05$ and $.01$ for $2n = 10, 20, 30, 40, 50$.

In order to indicate how to find moments of run variables let us consider the case of r_1. We shall first find the factorial moments $E(x^{(a)})$ where

$$x^{(a)} = x(x-1)(x-2)\ldots(x-a+1) = x!/(x-a)!,$$

for they are easier to find then ordinary moments in the present problem. From them the ordinary moments may be found since $E(x^{(i)})$ is a linear function of the first i ordinary moments. Letting $i = 1,2,\ldots,a$, we obtain a system of a linear equations which may be solved to obtain the ordinary moments as linear functions of the factorial moments.

We have

$$(k) \qquad E(r_1^{(a)}) = \sum_{r_1=1}^{n_1} r_1^{(a)} p(r_1) = \sum_{r_1=1}^{n_1} \frac{r_1!}{(r_1-a)!}\, p(r_1) .$$

In order to evaluate (k) we use the following identity:

$$(l) \qquad \sum_{i=0}^{B} \frac{A!}{(C+i)!(A-C-i)!} \cdot \frac{B!}{i!(B-i)!} = \frac{(A+B)!}{(C+B)!(A-C)!}$$

which follows at once by equating coefficients of x^C in the expansion of

$$(m) \qquad (1+x)^A(1+\tfrac{1}{x})^B = \frac{(1+x)^{A+B}}{x^B} .$$

Therefore we have upon substituting $p(r_1)$ from (f) into (k), simplifying, and using (l)

$$(n) \qquad E(r_1^{(a)}) = (n_2+1)^{(a)} \cdot \sum_{r_1} \frac{(n_1-1)!}{(r_1-1)!(n_1-r_1)!} \cdot \frac{(n_2+1-a)!}{(r_1-a)!(n_2+1-r_1)!} \Big/ \frac{n!}{n_1!n_2!}$$

$$= (n_2+1)^{(a)} \frac{(n-a)!}{(n_1-a)!n_2!} \Big/ \frac{n!}{n_1!n_2!} .$$

From this result we find

$$E(r_1) = \frac{(n_2+1)n_1}{n} ,$$

$$\sigma^2_{r_1} = \frac{(n_2+1)^{(2)}n_1^{(2)}}{n\, n^{(2)}} .$$

A similar expression holds for $E(r_2^{(a)})$.

If the a's and b's are regarded as elements in a sample of size n from a binomial population in which p and q represent the probabilities associated with a and b, respectively, then n_1, the number of a's, is a random variable distributed according to the binomial law ${}_nC_{n_1} p^{n_1} q^{n_2}$. The probability laws analagous to (b), (c), (d), (e), (f) when n_1 is regarded as a random variable in this manner are simply obtained by multiplying each of these probability laws by ${}_nC_{n_1} p^{n_1} q^{n_2}$.

10.12 Case of k Kinds of Elements

The theory of runs has been extended to the case of several kinds of elements by Mood*. If there are k kinds of elements, say $a_1, a_2, \ldots, a_k$, denote by r_{ij} the number of runs of a_i of length j. Let r_i be the total number of runs of a_i. Mood has shown that

*A. M. Mood, "The Theory of Runs", Annals of Math. Stat., Vol. XI (1940) .

the joint distribution law of the r_{ij} is given by

$$(a) \qquad p(r_{ij}) = \prod_{i=1}^{k} \frac{r_i!}{r_{i1}!r_{i2}!\ldots r_{in_i}!} \cdot F(r_1,r_2,\ldots,r_k) \Big/ \frac{n!}{n_1!n_2!\ldots n_k!},$$

where $F(r_1,r_2,\ldots,r_k)$ is the number of ways r_1 objects of one kind, r_2 objects of a second kind, and so on, can be arranged so that no two adjacent objects are of the same kind. $F(r_1,r_2,\ldots,r_k)$ is the coefficient of $x_1^{r_1}x_2^{r_2}\ldots x_k^{r_k}$ in the expansion of

$$(b) \qquad (x_1+x_2+\ldots+x_k)^{k}(x_2+x_3+\ldots+x_k)^{r_1-1}(x_1+x_3+\ldots+x_k)^{r_2-1}\ldots(x_1+x_2+\ldots+x_{k-1})^{r_k-1}.$$

The argument for establishing (a) is very similar to that for the case of $k = 2$ and will not be repeated. Mood showed that the joint distribution function of $r_1,r_2,\ldots,r_k$ is given by

$$(c) \qquad p(r_1,r_2,\ldots,r_k) = \prod_{i=1}^{k} \binom{n_1-1}{r_1-1} \cdot F(r_1,r_2,\ldots,r_k) \Big/ \frac{n!}{n_1!n_2!\ldots n_k!},$$

which we state without proof. Various moment formulas and asymptotic distribution functions have been derived by Mood in the paper cited.

If instead of holding $n_1,n_2,\ldots,n_k$ fixed in the run problem for k kinds of elements, we allow the n's to be random variables with probability function $\pi(n_1,n_2,\ldots,n_k)$ (e. g., the multinomial distribution with $\sum_1^k n_i=n$), the run distribution functions (a) and (c) would simply be multiplied by $\pi(n_1,n_2,\ldots,n_k)$.

10.2 Application of Run Theory to Ordering Within Samples

Suppose $O_{2n+1}(x_1,x_2,\ldots,x_{2n+1})$ is a sample from a population in which x is a continuous random variable. Let $\tilde{x}$ be the median value of x in the sample. Let each sample value of $x < \tilde{x}$ be called a and each sample value of $x > \tilde{x}$ be called b. There are n a's and n b's in the sample, ignoring the median (which is neither). Now suppose we consider all possible orders in which the sample x's could have been drawn (ignoring the median in each case). It is clear that all of the run distribution functions (b), (c), (d), (e), (f) are applicable, for $n_1 = n_2 = n$, to this aggregate of possible orders of the x's (i. e. a's and b's) in the sample. If there is an even number, say 2n, items in the sample, we can take any number between the two middle values of x in the sample as a number for dividing the x's into a's and b's, and our run theory is immediately applicable to this case with $n_1 = n_2 = n$. In general if in a sample of size $kn + k - 1$ we choose the $(n+1)$th, $(2n+2)$th, $(3n+3)$th,....,$(k-1)(n+1)$th values of x in increasing order of magnitude as points of division, and let all x's less than the $(n+1)$th x be denoted by a_1, those

between the (2n+2)th and ($\frac{n+1}{2}$)th by a_2, and so on, we then reduce our sample to n a_1's, n a_2's,.....,n a_k's. Ignoring the k - 1 x's used for division points, it is clear that run theory for k kinds of objects is applicable to the aggregate of all possible orders in which sample x's could occur (ignoring the x's used for division points). The points of division can, of course, be taken so as to yield an arbitrary number of a_1's, a_2's, etc.

By classifying the values of x in a sample into a's and b's (or more generally into a_1's, a_2's,...,a_k's) and using the theory of runs we have a basis for testing the hypothesis of randomness in the sample as far as order is concerned. The more commonly used tests of the hypothesis of randomness based on run theory are:

(1) Number of runs of a's, for which the distribution is (f), §10.11. For given values of n_1 and n_2, the test consists in finding the largest value of r_1 (the number of runs of a's), say r_1^o, for which $Pr(r_1 \leqq r_1^o) \leqq \epsilon$, e. g., for $\epsilon = .05$. A similar statement may be made concerning runs of b's.

(2) Total number of runs of a's and of b's having distribution (g), §10.11. Again, the test consists in finding the largest value of u, say u^o, for which $Pr(u \leqq u^o) \leqq \epsilon$, for given values of n_1 and n_2.

(3) At least one run of a's (or b's) of at least length s, for $n_1 = n_2 = n$, based on the distribution (i), §10.11. The test consists of finding the smallest value of s for which probability (i) is $\leqq \epsilon$.

(4) At least one run of either a's or b's of at least length s, for $n_1 = n_2 = n$, based on the distribution (j), §10.11. The test consists of finding the smallest s for which probability (j) is $\leqq \epsilon$.

The distribution theory of each of these tests has been determined under the assumption that the hypothesis of randomness is true, with a view to controlling only Type I (see §7.3) errors. Type II errors for these tests have never been investigated, i. e., probability theory of the tests when some alternative weighting scheme (other than equal weights) is used for the different possible arrangements of a's and b's.

It should be noted by the reader that the theory of runs developed in §10.11 is not applicable to the following type of problem of reducing a sample to two kinds of elements: Suppose $x_1, x_2, \ldots, x_n$ are elements of a sample from a population with a continuous distribution function. Consider an arbitrary order of these n x's, and between each successive pair of elements write a if the left number of the pair is smaller than the right and b if it is larger. We then have reduced the sample to n - 1 a's and b's. We may define runs of a's and b's as before, but the theory of arrangements of the a's and b's as defined from the corresponding arrangements, and hence the distribution theory of runs of this type, is an unsolved problem in combinatorial statistics.

10.3 Matching Theory

A problem which frequently arises in combinatorial statistics is one which may be conveniently described by an example of card matching. Suppose each of two decks of ordinary playing cards is shuffled and let a card be dealt from each deck. If the two cards are of the same suit let us call the result a match. Let this procedure be continued until the entire 52 pairs of cards are dealt. There will be a total number of matches, say h. Each possible permutation of one deck compared with each possible permutation of the second deck will yield a value of h between 0 and 52, inclusive. Therefore if we consider all of these possible permutations with equal weight, we inquire as to what will be the distribution function of h in this set of permutations. Similarly if we consider three decks D_1, D_2, and D_3 of cards to be shuffled and matched we would have triple matches and three varieties of double matches. A triple match would occur if the three cards in a single dealing from the three decks were of the same suit. As for double matches, they would occur between decks D_1, D_2, between D_1, D_3 and between D_2, D_3. The problem arises as to what will be the distribution of triple matches and of the three varieties of double matches.

Extensions of the problem to more than three decks, to decks with arbitrary numbers of cards in each suit and an arbitrary number of suits suggest themselves at once. In this section we shall present some techniques for dealing with this problem without attempting to be exhaustive. It will be convenient to continue our discussion in card terminology, for no particular advantage is gained in introducing more general terminology. The generality of the results for objects or elements other than cards is obvious.

10.31 Case of Two Decks of Cards

Suppose we have a deck D_1 of n cards, each card belonging to one and only one of the k suits $C_1, C_2, \ldots, C_k$. Let $n_{11}, n_{12}, \ldots, n_{1k}$ $(\sum_{i=1}^{k} n_{1i}=n)$ be the number of cards belonging to $C_1, C_2, \ldots, C_k$, respectively. Let D_2 be another deck of n cards, each card belonging to one and only one of the classes $C_1, C_2, \ldots, C_k$. Let $n_{21}, n_{22}, \ldots, n_{2k}$ $(\sum_{i=1}^{k} n_{2i}=n)$ be the number of cards in D_2 belonging to $C_1, C_2, \ldots, C_k$, respectively.

The problem is to determine the probability of obtaining h matches under the assumption of random pairing of the cards. In other words, we wish to find the number of ways the two decks of cards can be arranged so as to obtain exactly h matches. Dividing this number by N, the total number of ways the two decks can be arranged, we obtain the probability of obtaining h matches under random pairing. The value of N is simply the total number of ways the two suits can be permuted, and is given by the product of two multinomial coefficients:

(a) $$N = \frac{n!}{\prod_1^k n_{1i}!} \cdot \frac{n!}{\prod_1^k n_{2i}!} .$$

To determine N(h), consider the enumerating function*

(b) $$\phi = \left(\sum_{i,j=1}^{k} x_i y_j e^{\delta_{ij}\theta}\right)^n ,$$

where $\delta_{ij} = 1$, if $i = j$, and 0, if $i \neq j$. We associate the auxiliary variables $x_1, x_2, \ldots, x_k$ with the suits $C_1, C_2, \ldots, C_k$ respectively of the first deck, and the auxiliary variables $y_1, y_2, \ldots, y_k$ with the corresponding suits of the second deck. ϕ is the product of n identical expressions, each expression consisting of the sum of k^2 terms, each term being a product of an x and a y. The term $x_i y_j e^{\delta_{ij}\theta}$ in any one of the n factors corresponds to the event of a card in suit C_i of the first deck being paired against a card in suit C_j of the second deck. If $i = j$ we have a match, and e^θ occurs as a factor. Now suppose we pick a typical term in the product given in (b). Such a term would be of the form

(c) $$(x_{i_1} y_{j_1} e^{\delta_{i_1 j_1}\theta})(x_{i_2} y_{j_2} e^{\delta_{i_2 j_2}\theta}) \ldots (x_{i_n} y_{j_n} e^{\delta_{i_n j_n}\theta}).$$

This general term corresponds to the event of n pairings as follows: a pairing between C_{i_1} of D_1 and C_{j_1} of D_2; a pairing between C_{i_2} of D_1 and C_{j_2} of D_2; ....; and a pairing between C_{i_n} of D_1 and C_{j_n} of D_2. Now if the compositions of D_1 and D_2 are specified as $n_{11}, n_{12}, \ldots, n_{1k}$ and $n_{21}, n_{22}, \ldots, n_{2k}$, respectively, then it follows that the only terms in the expansion of (b) which have any meaning for pairings of these two decks of cards are those of the form

(d) $$e^{h\theta} \cdot x_1^{n_{11}} x_2^{n_{12}} \ldots x_k^{n_{1k}} \, y_1^{n_{21}} y_2^{n_{22}} \ldots y_k^{n_{2k}} ,$$

where h is an integer such that $0 \leq h \leq n$. It should be noted that such terms may not

---

*Various authors have considered various enumerating functions, but the one which we shall use was devised by I. L. Battin, "On the Problem of Multiple Matching", Annals of Math. Stat., Vol. XIII (1942). Battin's function is relatively easy to handle and has the advantage of representing the two decks of cards symmetrically in the notation and operations. It extends readily to the case of several decks of cards. The reader should refer to Battin's paper for a fairly extensive bibliography on the matching problem.

exist for some values of h between 0 and π, which means that it is not always possible to have any arbitrary number of matches for given deck compositions. The term given in (d) corresponds to some arrangement of the two decks of cards such that there are exactly h matches. In general there are many such terms. Therefore, if we expand ϕ and determine the coefficient of the expression given by (d) we obtain the value of N(h), the number of ways in which h matches can occur. To simplify our notation let $K_h(\phi)$ denote the operation of taking the coefficient of expression (d) in the expansion of ϕ. We may rewrite ϕ as

$$\text{(e)} \qquad \phi = \left\{(\sum_1^k x_1y_1)e^\theta + \left[(\sum_1^k x_1)(\sum_1^k y_1) - (\sum_1^k x_1y_1)\right]\right\}^n .$$

Expanding we have

$$\text{(f)} \qquad \phi = \sum_{h=0}^{n} \binom{n}{h} e^{h\theta}(\sum_1^k x_1y_1)^h \left[(\sum_1^k x_1)(\sum_1^k y_1) - (\sum_1^k x_1y_1)\right]^{n-h} .$$

Expanding the expression in [ ], we have

$$\text{(g)} \qquad [\ ]^{n-h} = \sum_{g=0}^{n-h} \binom{n-h}{g}(-1)^{n-g-h}(\sum_1^k x_1)^g(\sum_1^k y_1)^g(\sum_1^k x_1y_1)^{n-g-h}.$$

Inserting this expression into (f), and expanding $(\sum_1^k x_1)^g\ (\sum_1^k y_1)^g\ (\sum_1^k x_1y_1)^{n-g}$, we find

$$\text{(h)} \qquad N(h) = K_h(\phi) = \sum_{g=0}^{n-h}(-1)^{n-h-g}\binom{n}{h}\binom{n-h}{g}T_g ,$$

where

$$\text{(i)} \qquad T_g = \sum_{s_i} \frac{(g!)^2(n-g)!}{\prod_1^k [(n_{1i}-s_i)!(n_{2i}-s_i)!s_i!]}$$

the summation extending over all positive integral (or zero) values of the s_i such that $\sum_1^k s_i = n-g$ and $n_{1i}-s_i \geq 0$, $n_{2i}-s_i \geq 0$, $i = 1,2,\ldots,k$. The probability P(h) of obtaining h matches is therefore N(h)/N, where N is given by (a).

For the case $k = 2$, the probability of h matches reduces to the following expression

$$\text{(j)} \qquad P(h) = \frac{\binom{n}{h}\binom{h}{i}\binom{n-h}{j}}{\binom{n}{n_{11}}\binom{n}{n_{22}}} ,$$

where $i = \frac{1}{2}(n_{11}-n_{22}+h)$, $j = \frac{1}{2}(n_{11}+n_{22}-h)$. Unless h is such that for given values of n_{11} and n_{22}, $n_{11} \pm (n_{22}-h)$ are positive even integers or 0, then $P(h) = 0$.

Greville* has given the distribution of h in a slightly different form and by another method.

Moments of the random variable h can be found directly from the enumerating function ϕ. We have

$$E(h^p) = \frac{1}{N}\sum_{h=0}^{n}\frac{\partial^p}{\partial\theta^p}[K_h(\phi)]$$

$$= \frac{1}{N}\frac{\partial^p}{\partial\theta^p}[\sum_{h=0}^{n}K_h(\phi)] \tag{k}$$

$$= \text{coefficient of } x_1^{n_{11}}x_2^{n_{12}}\ldots x_k^{n_{1k}}y_1^{n_{21}}y_2^{n_{22}}\ldots y_k^{n_{2k}}$$

$$\text{in the expansion of } \frac{1}{N}[\frac{\partial^p\phi}{\partial\theta^p}]_{\theta=0}.$$

The reader will find it instructive to carry out this operation for p = 1,2, and find that

$$E(h) = \sum_1^k \frac{n_{11}n_{21}}{n},$$

$$\sigma_h^2 = E(h^2) - [E(h)]^2 \tag{l}$$

$$= \frac{1}{n^2(n-1)}[(\sum_1^k n_{11}n_{21})^2 - n\sum_1^k n_{11}n_{21}(n_{11}+n_{21}) + n^2\sum_1^k n_{11}n_{21}].$$

It should be noted that our results can be readily extended to the case of two decks of cards in which the total numbers of cards are different or where one or more of the suits may have no cards at all. To consider the case of unequal total numbers of cards, say n_1 in deck D_1 and n_2 in deck D_2 where, without loss of generality, we can let $n_1 > n_2$, we simply add to D_2 $n_1 - n_2$ dummy cards, and consider them as a new suit. We would thus have k + 1 suits of cards, where the (k+1)-th suit is empty in D_1, i. e. $n_{1k+1} = 0$, $n_{2k+1} = n_1 - n_2$. The procedure from here on is just as before. The case in which some of the suits are empty in one or both decks is taken into account by specifying the values of the corresponding n_{11} or n_{21} as 0 in expanding ϕ and collecting terms.

The reader should note that if a <u>score</u> s_{ij} is assigned to a pairing in which the D_1 card belongs to the i-th suit and the D_2 card belongs to the j-th suit, then one can find the distribution of the <u>total</u> score T in n pairings (i. e., when the two decks are paired against each other) under the assumption of random matching, by replacing δ_{ij} by

*T. N. E. Greville, "The Frequency Distribution of a General Matching Problem", <u>Annals of Math. Stat.</u>, Vol. XII (1941).

s_{ij} in (b) and finding the coefficient of

$$\frac{1}{N} e^{T\theta} x_1^{n_{11}} x_2^{n_{21}} \ldots x_k^{n_{1k}} y_1^{n_{21}} y_2^{n_{22}} \ldots y_k^{n_{2k}}$$

in the expansion of the resulting expression. The procedure for finding E(T) and σ_T^2 and higher moments is the same as that for dealing with the moments of h with s_{ij} substituted for δ_{ij}.

10.32 Case of Three or More Decks of Cards

Suppose we have a third deck of cards, say D_3. Let the numbers of cards belonging to suits $C_1, C_2, \ldots, C_k$ be $n_{31}, n_{32}, \ldots, n_{3k}$. A triple match has been defined as one in which the triplet of cards (one from each deck) are of the same suit. A double match between D_1 and D_2 will occur when the cards from D_1 and D_2 in a triplet are of the same suit but different from the suit of the card from D_3 in the triplet. Double matches between D_1, D_3 and D_2, D_3 are similarly defined. If in the complete set of n triplets from the three decks we let h_{123} be the number of triple matches, h_{12} the number of double matches between D_1, D_2, with similar meanings for h_{13} and h_{23}, we may obtain the distributions and moments of the h's from the following enumerating function:

$$\text{(a)} \qquad \phi = \Big(\sum_{i,j,k=1}^{n} x_i y_j z_k e^{\delta_{ijk}\theta_{123} + \delta_{ij}\theta_{12} + \delta_{ik}\theta_{13} + \delta_{jk}\theta_{23}} \Big)^n,$$

where $\delta_{ijk} = 1$, if $i = j = k$, and 0 otherwise. The remaining δ's are defined as for the 2-deck problem. By following an argument similar to that for the 2-deck problem, it will be noted that the number of ways in which the three decks of cards can be permuted so as to obtain h_{123} triple matches, and h_{12}, h_{13}, h_{23} double matches between D_1, D_2; D_1, D_3; and D_2, D_3, respectively, is given by the coefficient of

$$\text{(b)} \qquad e^{h_{123}\theta_{123} + h_{12}\theta_{12} + h_{13}\theta_{13} + h_{23}\theta_{23}} \cdot Q,$$

where

$$Q = x_1^{n_{11}} x_2^{n_{12}} \ldots x_k^{n_{1k}} \cdot y_1^{n_{21}} y_2^{n_{22}} \ldots y_k^{n_{2k}} \cdot z_1^{n_{31}} z_2^{n_{32}} \ldots z_k^{n_{3k}}$$

in the expansion of ϕ.

This coefficient and hence the joint probability law of the h's is rather cumbersome and will not be given here. As in the case of the 2-deck problem we may find moments and joint moments of the h's by performing differentiations on ϕ with respect to the θ's, that is,

$$(c)\qquad E(h_{123}^{r_1}h_{12}^{r_2}h_{13}^{r_3}h_{23}^{r_4}) = \frac{1}{N}\left\{\text{Coeff. of } Q \text{ in } \left[\frac{\partial^{r_1+r_2+r_3+r_4}\phi}{\partial\theta_{123}^{r_1}\partial\theta_{12}^{r_2}\partial\theta_{13}^{r_3}\partial\theta_{23}^{r_4}}\right]_{\theta\text{'s}=0}\right\},$$

where

$$(d)\qquad N = \frac{(n!)^3}{\prod_1^k (n_{11}!n_{21}!n_{31}!)}.$$

The mean values of the h's are the following:

$$E(h_{123}) = \sum_{i=1}^{k}\frac{n_{11}n_{21}n_{31}}{n^2},$$

$$E(h_{12}) = \frac{1}{n^2}\left(\sum_1^k n_{11}n_{21}\right)\left(\sum_1^k n_{31}\right),$$

with similar expressions for $E(h_{13})$ and $E(h_{23})$. The reader may refer to Battin's paper for second moments.

The extension of our technique to the problem of determining the distribution and moments of the numbers of hits for various orders of multiple matching when more than three decks of cards are involved is immediate. The extension of the results to the case of decks of unequal numbers of cards, empty suits, etc., when three or more decks are considered, is straightforward.

10.4 Independence in Contingency Tables.

In this section we shall consider the problem of testing the independence of a two-way classification on basis of a sample of n elements, each element belonging to one and only one the classes $A_1, A_2, \ldots, A_r$ and to one and only one of the classes $B_1, B_2, \ldots, B_s$. In the sample, let n_{ij} be the number of elements belonging to A_i and B_j. Let $\sum_{j=1}^{s} n_{ij} = n_{i.}$, $\sum_{i=1}^{r} n_{ij} = n_{.j}$, $\sum_{i,j} n_{ij} = n$. The number of elements belonging to A_i is $n_{i.}$ and the number to B_j is $n_{.j}$. The problem is to test the hypothesis of the independence of the A and B classification. We shall consider two approaches to this problem. The first (§10.41) is a pure combinatorial approach based on partition theory in which the set of all possible partitions of n into rs components n_{ij} satisfying the marginal conditions listed above are investigated. The second approach (§10.42), which is Karl Pearson's original treatment of the problem, is an application of the theory of sampling from a multinomial population consisting of the rs classes (A_iB_j) $i = 1,2,..,r$; $j = 1,2,..,s$.

10.41 The Partitional Approach

In this section we shall consider the problem of determining the number of ways of partitioning the integer n into rs integers (or zero) n_{ij} $(i=1,2,\ldots,r;\ j=1,2,\ldots,s)$

such that $\sum_{j=1}^{s} n_{ij} = n_{i.}$ and $\sum_{i=1}^{r} n_{ij} = n_{.j}$ are fixed. The technique discussed in §10.3 can be extended so as to accomplish this enumeration. We shall then find the mean values of certain functions of the n_{ij} over this set of partitions.

We may represent the n_{ij}, $n_{i.}$, $n_{.j}$ and n in the following contingency table:

(a)

					Total
	n_{11}	n_{12}	. . . .	n_{1s}	$n_{1.}$
	n_{21}	n_{22}	. . . .	n_{2s}	$n_{2.}$
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	n_{r1}	n_{r2}	. . . .	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	. . . .	$n_{.s}$	n .

Consider the enumerating function

$$\phi = \prod_{i=1}^{r} \Big(\sum_{j=1}^{s} x_j e^{\theta_{ij}} \Big)^{n_{i.}} , \tag{b}$$

which is the product of n factors, $n_{1.}$ of which are $\sum_{1}^{s} x_j e^{\theta_{1j}}$, $n_{2.}$ of which are $\sum_{1}^{s} x_j e^{\theta_{2j}}$ and so on. A typical term in the expansion of this product of n factors is of the form

$$\Big(x_1^{\sum_1 n_{11}} e^{\sum_1 n_{11}\theta_{11}} \Big) \cdot \Big(x_2^{\sum_1 n_{12}} e^{\sum_1 n_{12}\theta_{12}} \Big) \ldots \Big(x_s^{\sum_1 n_{1s}} e^{\sum_1 n_{1s}\theta_{1s}} \Big), \tag{c}$$

where n_{11} is the number of times $x_1 e^{\theta_{11}}$ is taken from the $n_{1.}$ factors $\big(\sum_j x_j e^{\theta_{1j}}\big)^{n_{1.}}$, with similar meanings for $n_{12}, n_{13}, \ldots, n_{1s}$. If $\sum_1 n_{11} = n_{.1}$, $\sum_1 n_{12} = n_{.2}, \ldots$, then (c) corresponds to one way of partitioning n into the set n_{ij} so that $\sum_1 n_{ij} = n_{.j}$ and $\sum_j n_{ij} = n_{i.}$. To find the total number of ways of partitioning n into the given set n_{ij} we must determine how many individual terms in the expansion of (b) are identical with (c). In other words we are to find the coefficient of

$$\Big(\prod_j x_j^{n_{.j}} \Big) e^{\sum_{i,j} n_{ij}\theta_{ij}} \tag{d}$$

in the expansion of (b).

Expanding each of the terms $\big(\sum_{j=1}^{s} x_j e^{\theta_{ij}}\big)^{n_{i.}}$, i = 1,2,..,r, by the multinomial law and multiplying the results and taking the coefficient of the expression (d), we find at once that the number of partitions of n into the sets of values n_{ij}, subject to the marginal conditions $\sum_1 n_{ij} = n_{.j}$, $\sum_j n_{ij} = n_{i.}$, is

(e) $$\prod_i \left(\frac{n_{i.}!}{\prod_j n_{ij}!}\right) .$$

The total number of ways of partitioning n, subject to the marginal conditions mentioned above, is

(f) $$\frac{n!}{\prod_j n_{.j}!} .$$

Therefore the probability of partitioning n into the particular set of values n_{ij}, assuming all ways of making partitions (subject to the marginal conditions) equally likely, is given by the ratio of expression (e) to expression (f).

The moments of the n_{ij} may be found directly from the probability law of the n_{ij}. Consider first the problem of determining the h-th factorial moment of a particular n_{ij}, say $n_{\alpha\beta}$. We have

(g) $$E(n_{\alpha\beta}^{(h)}) = \sum n_{\alpha\beta}^{(h)} \prod_i \left(\frac{n_{i.}!}{\prod_j n_{ij}!}\right) \Big/ \frac{n!}{\prod_j n_{.j}!} ,$$

where $\sum$ denotes summation over all values of the n_{ij} subject to the usual marginal conditions. Now when h = 0, we know that the right hand side of (g) is simply the sum of the probability function of the n_{ij} over all possible values of the n_{ij} and is therefore unity, which amounts to the statement that

(h) $$\sum \prod_i \left(\frac{n_{i.}!}{\prod_j n_{ij}!}\right) = \frac{n!}{\prod_j n_{.j}!} .$$

Now the numerator on the right hand side of (g) may be written as

(i) $$n_{\alpha.}^{(h)} \sum \prod_i \left(\frac{n'_{i.}!}{\prod_j n'_{ij}!}\right),$$

where $n'_{i.} = n_{i.}$ for all i except $i = \alpha$ and $n'_{\alpha.} = (n_{\alpha.} - h)$ and $n'_{ij} = n_{ij}$ for all i,j except for $i = \alpha$, $j = \beta$ and $n'_{\alpha\beta} = n_{\alpha\beta} - h$. Now perform the summation indicated in (i) over all values of the n'_{ij} subject to the conditions $\sum_j n'_{ij} = n'_{i.}$ and $\sum_i n'_{ij} = n'_{.j}$ where $n'_{.j} = n_{.j}$ except when $j = \beta$ and $n'_{.\beta} = n_{.\beta} - h$. It follows from (h) that the value of this sum is

$$\frac{(n-h)!}{\prod_j n'_{.j}!} .$$

Therefore we have

(j) $$E(n_{\alpha\beta}^{(h)}) = n_{\alpha.}^{(h)} \frac{(n-h)!}{\prod_j n_{.j}!} \Big/ \frac{n!}{\prod_j n_{.j}!} = \frac{n_{\alpha.}^{(h)} n_{.\beta}^{(h)}}{n^{(h)}} .$$

It is clear that h must be less than each of the numbers $n_{\alpha.}$ and $n_{.\beta}$. For h = 1 and 2 we have

(k) $$E(n_{\alpha\beta}) = \frac{n_{\alpha.} n_{.\beta}}{n}$$

$$E(n_{\alpha\beta}^2 - n_{\alpha\beta}) = \frac{n_{\alpha.}^{(2)} n_{.\beta}^{(2)}}{n^{(2)}} .$$

Hence

(l) $$E(n_{\alpha\beta}^2) = \frac{n_{\alpha.}^{(2)} n_{.\beta}^{(2)}}{n^{(2)}} + \frac{n_{\alpha.} n_{.\beta}}{n} .$$

By a similar argument one can find joint factorial moments of two or more of the n_{ij}. For example,

(m) $$E(n_{\alpha\beta}^{(h)} \cdot n_{\gamma\delta}^{(g)}) = \frac{n_{\alpha.}^{(h)} n_{.\beta}^{(h)} n_{\gamma.}^{(g)} n_{.\delta}^{(g)}}{n^{(g+h)}} , \qquad \alpha \neq \gamma, \ \beta \neq \delta$$

$$= \frac{n_{\alpha.}^{(g+h)} n_{.\beta}^{(h)} n_{.\delta}^{(g)}}{n^{(g+h)}} \qquad \alpha = \gamma, \ \beta \neq \delta .$$

A similar expression holding for $\alpha \neq \gamma$, $\beta = \delta$. The restrictions on the size of g and h are obvious. These moments can also be found directly from the enumerating function ϕ by carrying out appropriate differentiations on the θ_{ij} then setting the θ's = 0 and collecting appropriate coefficients.

The criterion which Karl Pearson def'ned for testing the hypothesis of row-column independence in r by s contingency tables is defined as follows

(n) $$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}} ,$$

which is a quadratic form in the n_{ij}. It should be noted that χ^2 is simply the sum of the squared differences between each n_{ij} and its mean value (under the assumption of independence or "randomness"), each squared difference weighted inversely by the mean value of the corresponding n_{ij}. This inverse weighting scheme suggests itself fairly readily in the Pearson approach to be considered in §10.42. The mean value of χ^2 may easily be found by making use of formulas (k) and (l), and is

(o) $$E(\chi^2) = \frac{n}{n-1}(r-1)(s-1).$$

By using formulas (j) and (m) for the appropriate values of g and h, the variance and higher moments of χ^2 may be found.

10.42 Karl Pearson's Original Chi-Square Problems and its Application to Contingency Tables.

Suppose $\prod$ is a multinomial population in which each element belongs to one and only one of the classes $C_1, C_2, \ldots, C_k$. Let $p_1, p_2, \ldots, p_k$ $(\sum_1^k p_i = 1)$ be the probabilities associated with $C_1, C_2, \ldots, C_k$ respectively. In a sample of size n let $n_1, n_2, \ldots, n_k$ be the numbers of elements falling into $C_1, C_2, \ldots, C_k$ respectively. We have seen (§3.12) that the probability law of the n_i is

(a) $$\frac{n!}{n_1!n_2!\ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}.$$

It was shown in §3.12 that $E(n_i) = np_i$. In view of the Central Limit Theorem (§4.21) it is clear that the limiting distribution as $n \to \infty$ of each of the quantities

$$\frac{(\frac{n_i}{n} - p_i)\sqrt{n}}{\sqrt{p_i(1-p_i)}} \qquad i = 1,2,\ldots,k$$

is N(0,1). Now let us investigate the limiting joint distribution of the set

$$x_i \equiv \frac{(n_i - np_i)}{\sqrt{n}} \qquad i = 1,2,\ldots,k.$$

Since $\sum_1^k x_i \equiv 0$ only k - 1 of the x_i are functionally independent. It is sufficient to consider the limiting joint distribution of the first k - 1 of the x_i. The m. g. f. of $x_1, x_2, \ldots, x_{k-1}$ is

(b) $$\phi = E(e^{\sum_1^{k-1} \theta_i x_i}) = \sum (e^{\sum_1^{k-1} \theta_i (n_i - np_i)/\sqrt{n}} \frac{n!}{\prod_1 n_i!} \prod_1 p_i^{n_i})$$

$$= e^{-\sqrt{n}\sum_1^{k-1} \theta_i p_i} (p_1 e^{\theta_1/\sqrt{n}} + p_2 e^{\theta_2/\sqrt{n}} + \ldots + p_{k-1} e^{\theta_{k-1}/\sqrt{n}} + p_k)^n.$$

Expanding each of the exponentials in ( ) and taking logarithms, we have

$$\log \phi = -\sqrt{n}\sum_{1}^{k-1}\theta_i p_i + n \log\left(1 + \sum_{1}^{k-1}\frac{\theta_i p_i}{\sqrt{n}} + \sum_{1}^{k-1}\frac{\theta_i^2 p_i}{2n} + \ldots\right)$$

(c)
$$= \sum_{1}^{k-1}\frac{\theta_i^2 p_i}{2} - \sum_{i,j=1}^{k-1}\frac{\theta_i\theta_j p_i p_j}{2} + O\left(\frac{1}{\sqrt{n}}\right).$$

Therefore we have

(d)
$$\lim_{n\to\infty} \phi = e^{\frac{1}{2}\sum_{i,j=1}^{k-1} A^{ij}\theta_i\theta_j},$$

where $A^{ij} = p_i\delta_{ij} - p_i p_j$, $i,j = 1,2,\ldots,k-1$, where $\delta_{ij} = 1$, $i = j$, and 0, $i \neq j$. Making use of the multivariate analogue of Theorem (C), §2.81, it follows that the limiting probability element for the distribution of the x_i is

(e)
$$\frac{\sqrt{A}}{(2\pi)^{\frac{k-1}{2}}} e^{-\frac{1}{2}\sum_{i,j=1}^{k-1} A_{ij}x_i x_j} dx_1\ldots dx_{k-1},$$

where $||A_{ij}|| = ||A^{ij}||^{-1}$. It may be readily verified by the reader that

(f)
$$A_{ij} = \frac{\delta_{ij}}{p_i} + \frac{1}{p_k},$$

and hence

(g)
$$\sum_{i,j=1}^{k-1} A_{ij}x_i x_j = \sum_{1}^{k-1}\frac{x_i^2}{p_i} + \frac{1}{p_k}\left(\sum_{1}^{k-1} x_i\right)^2.$$

We have seen, (§5.22), that if $x_1, x_2, \ldots, x_{k-1}$ are random variables having distribution (e) then $\sum_{i,j=1}^{k-1} A_{ij}x_i x_j$ is distributed according to a χ^2-law with $k-1$ degrees of freedom.

Now if we replace x_i by $(n_i - np_i)/\sqrt{n}$ in (g) denoting the result by χ^2, we obtain

(h)
$$\chi^2 = \sum_{1}^{k}\frac{(n_i - np_i)^2}{np_i}.$$

We conclude that the limiting distribution of χ^2 is identical with the distribution of $\sum_{1}^{k-1} A_{ij}x_i x_j$ where the x_i are distributed according to (e); that is to say, the limiting distribution of the expression in (h) is the χ^2-law with $k-1$ degrees of freedom. A rigorous proof of this statement is beyond the scope of this course, but it is a consequence of the following theorem which will be stated without proof:

Theorem (A): Let $x_1^{(n)},x_2^{(n)},\ldots,x_r^{(n)}$ be random variables having a joint c. d. f. for each n greater than some n_0. Let the limiting joint c. d. f. as $n \to \infty$ be $F(x_1,x_2,\ldots,x_r)$. Let $g(x_1,x_2,\ldots,x_r)$ be a (Borel measurable) function of $x_1,x_2,\ldots,x_r$. Then the limiting c. d. f. say $F(g)$ of $g(x_1^{(n)},x_2^{(n)},\ldots,x_r^{(n)})$ as $n \to \infty$ is given by

$$F(g) = \int_R dF(x_1,x_2,\ldots,x_r),$$

where R is the region in the x space for which $g(x_1,x_2,\ldots,x_r) \leqq g$.

We may summarize our results therefore in the following theorem, which is, in fact, a corollary of Theorem (A):

Theorem (B): Let O_n be a sample of size n from the multinomial distribution (a). Then the limiting distribution of $\sum_1^k \frac{(n_i-np_i)^2}{np_i}$ as $n \to \infty$ is the χ^2-distribution with k - 1 degrees of freedom.

Now let us consider the contingency problem described in the introduction of §10.4. In this case the multinomial population consists of rs classes (A_iB_j) i = 1,2,..r; j = 1,2,..,s. Let the probability associated with (A_iB_j) be $p_{ij}(\sum_{i,j} p_{ij}=1)$. It follows at once by Theorem (B) that

$$\sum_{i,j}(n_{ij}-np_{ij})^2/np_{ij} \tag{i}$$

has as its limiting distribution for $n \to \infty$, the χ^2-law with rs - 1 degrees of freedom. If the p_{ij} were known a priori, then the test given in (i) could be used for testing the hypothesis that the sample originated from a multinomial distribution having these values of p_{ij}. If the A and B classifications are independent in the probability sense then $p_{ij} = p_iq_j$ $(\sum_1^r p_i=1; \sum_1^s q_j=1)$. If the p_i and q_j were known a priori then (i) with $p_{ij} = p_iq_j$ can, of course, be used to test the hypothesis that the sample came from a multinomial population with probabilities p_iq_j.

But suppose neither the p_i nor q_j are known a priori, and that we wish merely to test the hypothesis of independence of the A and B classifications. Karl Pearson proposed the following test for this hypothesis

$$\chi_c^2 = \sum_{i,j}(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2 / \frac{n_{i.}n_{.j}}{n}, \tag{j}$$

where the $n_{i.}$ and $n_{.j}$ are defined in §10.41.

If we let

$$x_{ij} = \frac{n_{ij} - np_iq_j}{\sqrt{n}},$$

and express the n_{ij} in (j) in terms the x_{ij} we obtain

$$(k) \qquad \chi_c^2 = \sum_{i,j}(x_{ij}-x_{i.}q_j-x_{.j}p_i)^2/p_iq_j + O(\tfrac{1}{\sqrt{n}}),$$

where $x_{i.} = \sum_j x_{ij}$ and $x_{.j} = \sum_i x_{ij}$.

By following an argument similar to that used in determining the limiting distribution (e) of the x_i, $i = 1,2,\ldots,k-1$, we may find the limiting distribution of the x_{ij} (all i,j except i=r, j=s) to be normal multivariate. From this limiting distribution one finds that the distribution of $\sum_{i,j}(x_{ij}-x_{i.}q_j-x_{.j}p_i)^2/p_iq_j$ is the χ^2 distribution with (r-1)(s-1) degrees of freedom. By an argument similar to that embodied in Theorem (A) we may make the following statement:

Theorem (C): Let O_n be a sample from a multinomial population with the mutually exclusive classes (A_iB_j) $i = 1,2,\ldots,r$; $j = 1,2,\ldots,s$, in which the probability associated with (A_iB_j) is p_iq_j. Let χ_c^2 be defined as in (j). Then the limiting distribution of χ_c^2 as $n \longrightarrow \infty$ is the χ^2-distribution with (r-1)(s-1) degrees of freedom.

The reader may verify that the likelihood ratio criterion for testing the hypothesis specified by

$$\Omega: \quad p_{ij} > 0, \ \sum_{i,j} p_{ij} = 1$$

$$\omega: \quad p_{ij} = p_iq_j, \ \sum_i p_i = \sum_j q_j = 1,$$

that is, the hypothesis that the A and B classifications are independent is given by

$$\lambda = \frac{n^n(\prod_{i,j} n_{ij}^{n_{ij}})}{(\prod_i n_{i.}^{n_{i.}})(\prod_j n_{.j}^{n_{.j}})}.$$

It follows from Theorem (A), §7.2, that when the hypothesis of independence is true, the limiting distribution of $-2 \log \lambda$ is the χ^2-distribution with (r-1)(s-1) degrees of freedom.

10.5 Sampling Inspection

In a mass production process, suppose articles are produced in lots of N articles each, and suppose each article, upon inspection, can be classified as defective or non-defective. It is often uneconomical to carry out a program of 100% inspection. As an alternative, sampling methods of inspection applicable to each lot have been developed which have the property of guaranteeing that the percentage of defectives remaining

after applying the sampling inspection procedure in the long run (i. e. to a large number of lots) is not more than some preassigned value. Such sampling methods have been developed and put into operation by Dodge and Romig* of the Bell Telephone Laboratories. It should be pointed out that these sampling methods are essentially screening devices for reducing defectives after production, and are not devices for removing the causes of defectives. Methods for detecting the existence of causes of such defectives must be introduced further back into the production operations. In particular, statistical quality control methods** originally introduced by Shewhart, have been found useful in connection with this problem.

The mathematical problem involved in sampling inspection is one in combinatorial statistics. Dodge and Romig have developed two types of inspection sampling, single sampling and double sampling, which will be considered in turn. From a mathematical point of view, many sampling inspection schemes can be devised which guarantee quality of outgoing products in the sense mentioned above.

10.51 Single Sampling Inspection

Let p be the fraction of defectives in a lot L_N of size N. The number of defectives will be pN. Now let a sample O_n of size n be drawn from L_N. Giving all possible samples of size n equal weight, the probability of obtaining m defectives (and n - m non-defectives or conforming articles) in O_n is

$$(a) \qquad P_{m,n;pN,N} = \frac{\binom{N-pN}{n-m} \cdot \binom{pN}{m}}{\binom{N}{n}}, \qquad m=0,1,2,\ldots,r$$

where r is the smaller of n and Np. Let

$$(b) \qquad F(c;p,N,n) = Pr(m \leq c) = \sum_{m=0}^{c} P_{m,n;pN,n}.$$

It is easy to verify that if any two values of p and p' (pN and p'N being integers) are such that $p < p'$ then

$$(c) \qquad F(c;p,N,n) > F(c;p',N,n).$$

*H. F. Dodge and H. G. Romig "A Method of Sampling Inspection", Bell System Technical Journal, Vol. VIII (1929) and "Single Sampling and Double Sampling Inspection Tables", Bell System Technical Journal, Vol. XX (1941).

**See "Guide for Quality Control and Control Chart Method of Analyzing Data" (1941) and "Control Chart Method of Controlling Quality During Production" (1942), American Standards Association, New York.

Let p_t be the lot tolerance fraction defective, i. e. the maximum allowable fraction defective in a lot, which is arbitrarily chosen in advance (e. g., .01 or .05). Let

$$P_C = F(c;p_t,N,n).$$

P_C is known as the consumer's risk; it is (approximately) the probability that a lot with lot tolerance fraction defective p_t will be accepted without 100% inspection. It follows from (c) that if the lot fraction defective p exceeds p_t then the probability of accepting such a lot on basis of the sample is less than the consumer's risk. The probability of subjecting a lot with fraction defective actually equal to $\bar{p}$ (process average) to 100% inspection is

(d) $$P_P = 1 - F(c;\bar{p},N,n),$$

which is called producer's risk. It will be noted from (c) that the smaller the value of $\bar{p}$, the smaller will be the producer's risk.

The reader should observe that producer risk and consumer risk are highly analogous to Type I and Type II errors, respectively, (see §7.3) in the theory of testing statistical hypotheses as developed by Neyman and Pearson. In fact, historically speaking, the concept of producer and consumer risks in sampling inspection may be considered as the forerunner of the concept of Type I and Type II errors in the theory of testing statistical hypotheses.

Now suppose we make the following rules of action with reference to a sampled lot where c is chosen for given values of P_C, p_t, N, n:

(1) Inspect a sample of n articles.

(2) If the number of defectives in the sample does not exceed c, accept the lot.

(3) If the number of defectives in the sample exceeds c, inspect the remainder of the lot.

(4) Replace all defectives found by conforming articles.

Now let us consider the problem of determining the mean value of the fraction defectives remaining in a lot having fraction defective = p, after applying rules of action (1) to (4).

The probability of obtaining m defectives in a sample of size n is given by (a). If these m defectives are replaced by conforming articles and the sample is returned to the lot, the lot will contain pN - m defectives. Hence the probability of accepting a lot with pN - m defectives is given by (a), m = 0,1,2,...,c. The probability of inspecting the lot 100% is 1 - F(c;p,N,n), which, of course, is the probability of accepting a

lot with no defectives. Therefore the mean value of the fraction of defectives remaining after applying rules (1) to (4) is

$$\text{(e)} \qquad \tilde{p} = \sum_{m=0}^{c} \left(\frac{pN-m}{N}\right) P_{m,n;pN,N}.$$

The statistical interpretation of (e) is as follows: If a large number of lots each with fraction defective p are inspected according to rules (1) to (4), then the average fraction defective in all of these lots after inspection is $\tilde{p}$. For given values of c, n, and N, $\tilde{p}$ is a function of p, defined for those values of p for which Np is an integer, which has a maximum with respect to p. Denoting this maximum by $\tilde{p}_L$, it is called average outgoing quality limit. It can be shown that the larger the value of p beyond the value maximizing p, the smaller will be the value of $\tilde{p}$. The reason for this, of course, is that the greater the value of p, the greater the probability that each lot will have to be inspected 100%. If the consumer risk, n, and N are chosen in advance, then, of course, c and hence $\tilde{p}_L$ is determined. Thus, we are able to make the following statistical interpretation of these results:

If rules (1), (2), (3) and (4) are followed for lot after lot and for given values of c, n, N, the average fraction defective per lot after inspection never exceeds $\tilde{p}_L$, no matter what fractions defective exist in the lots before the inspection.

It is clear that there are various combinations of values of c and n, each having a $\tilde{p}$ with maximum $\tilde{p}_L$ (approximately) with respect to p.

The mean value of the number of articles inspected per lot for lots having fraction defective p is given by

$$\text{(f)} \qquad I = n + (N-n)(1-F(c;p,N,n)),$$

since n (the number in the sample) will be inspected in every lot and N - n (the remainder in the lot) will be inspected if the number of defectives in the sample exceeds c.

Thus, we have two methods of specifying consumer protection: (i) Lot quality protection obtained by specifying lot tolerance fraction defective p_t and consumer's risk P_C; (ii) Average quality protection in which average outgoing quality limit $\tilde{p}_L$ is specified.

By considering the various combinations of values of c and n corresponding to a given consumer's risk (or to a given average outgoing quality limit) there is, in general, a unique combination, for a given p and N, for which I is smaller than for any other. Such a combination of values of n and c together with a value of p as near to its actual value $\bar{p}$ in the incoming lots as one can "obtain" is, from a practical point of view, the

combination to use since amount of inspection is reduced to a minimum.

Extensive tabulations of pairs of values of c and n, for consumer's risk = 0.10, for values of N from 1 to 100,000,for lot tolerance fraction defective from .005 to .10, and for process average from .00005 to .05, all of the variables broken down into suitable groupings, have been prepared by Dodge and Romig. They have also made tabulations of pairs of values of c and n for given values of outgoing quality limit $\tilde{p}_L$ from .001 to .10, for values of N from 1 to 100,000 and for values of process average from .00002 to .10. Numerous approximations have been made to formulas (a), (b), (d), (e) and (f) for computation purposes, which the reader may refer to in the papers cited. For example, it is easy to verify that the Poisson law $e^{-pn}(pn)^m/m!$ is a good approximation to (a) if p and $\frac{n}{N}$ are both small, say <0.10.

10.52 Double Sampling Inspection

In double sampling inspection from a given lot of size N, the procedure for taking action regarding a given lot is as follows:

(1) A first sample of size n_1 is drawn from the lot.

(2) If the number of defectives is $\leq c_1$, the lot is accepted without further sampling.

(3) If the number of defectives in the first sample exceeds c_2 inspect the remainder of the lot.

(4) If the number of defectives in the first sample exceeds c_1 but not c_2, inspect a second sample of n_2 pieces.

(5) If the total number of defectives in both samples does not exceed c_2, accept the lot.

(6) If the total number of defectives in both samples exceeds c_2, inspect the remainder of the lot.

(7) Replace all defectives found by conforming articles.

As in the case of single sampling, we have two kinds of consumer protection: (i) Lot quality protection, and (ii) Average quality protection.

Consumer risk, the probability of accepting a lot with fraction defective p_t without 100 % inspection, is given by

$$(a) \qquad P_C = \sum_{m=0}^{c_1} P_{m,n_1;p_tN,N} + \sum_{i=1}^{c_2-c_1} \sum_{m=0}^{c_2-c_1-i} (P_{c_1+i,n_1;p_tN,N})(P_{m,n_2;p_tN-c_1-i,N-n_1}) .$$

The single sum in this formula is simply the probability of accepting the lot on basis of the first sample (i. e. Step (2)) and the double sum is the probability of accepting the

lot on basis of the first and second samples combined (i. e. Step (5)), after having failed to accept on basis of the first sample alone.

The mean value of the fraction of defectives per lot remaining after the defectives have been removed by the double sampling procedure, for lots having fraction defective p originally, is given by

$$\text{(b)} \qquad \tilde{p} = \sum_{m=0}^{c_1} \left(\frac{Np-m}{N}\right) P_{m,n_1;pN,N} + \sum_{i=1}^{c_2-c_1} \sum_{m=0}^{c_2-c_1-i} \left(\frac{Np-(c_1+i+m)}{N}\right)(P_{c_1+i,n_1;pN,N})(P_{m,n_2;pN-c_1-i,N-n_1}).$$

The mean value of the number of articles inspected per lot for lots having fraction defective p is

$$\text{(c)} \qquad I = n_1 + n_2\left(1-\sum_{m=0}^{c_1} P_{m,n_1;pN,N}\right) + (N-n_1-n_2)(1-P_a),$$

where P_a is the value of the probability given in (a) with p_t replaced by p.

For given values of N_1, n_1, n_2, c_1, c_2, it is clear that $\tilde{p}$ is a function of p, defined for those values of p for which Np is an integer, and has a maximum value $\tilde{p}_L$, the average outgoing quality limit. For a given value of N there are many values of n_1, n_2, c_1, c_2 which will yield the same value of $\tilde{p}_L$ (approximately), or will yield the same consumer risk (approximately) for a given lot tolerance fraction defective. Dodge and Romig have arbitrarily chosen as the basis for the relationship between n's and the c's the following rule: To determine n_1 and n_2 such that for given values of c_1 and c_2, n_1 and c_1 (as sample size and allowable defect number) provides the same consumer risk (approximately) as $n_1 + n_2$ and c_2 (as sample size and allowable defect number). The sense in which "approximately" is used is due to nearest integer restrictions. Even after this restriction there is enough choice left for combinations of n_1, n_2, c_1, c_2 to minimize I as given by (c). To determine the n's and c's under these conditions for given N, for given consumer risk, (or average outgoing quality) involves a considerable amount of computation. Dodge and Romig have prepared tables for double sampling analogous to those described at the end of §10.51 for single sampling.

For a given amount of consumer protection, a smaller average amount of inspection is required under doubling sampling than under single sampling, particularly for large lots and low process average fraction defective $\bar{p}$.

CHAPTER XI

AN INTRODUCTION TO MULTIVARIATE STATISTICAL ANALYSIS

A considerable amount of work has been done in recent years in the theory of sampling from normal multivariate populations and in the theory of testing statistical hypotheses relating to normal multivariate distributions. The two basic distribution functions underlying all of this work are the sample mean distribution (e) in §5.12, and the Wishart distribution, (k) in §5.6, of the second order sample moments. We have given a derivation of the distribution of means (§5.12) and a derivation of the Wishart distribution for the case of samples from a bivariate normal population, (§5.12). The general Wishart distribution was given in §5.5, without proof.

In the present chapter we shall present a geometric derivation* of the Wishart distribution, and consider applications of this distribution in deriving sampling distributions of several multivariate statistical functions and test criteria. The few sections which follow must be considered merely as an introduction to normal multivariate statistical theory. The reader interested in further material in this field is referred to the Bibliography for supplementary reading.

11.1 The Wishart Distribution

In §5.5, we presented a derivation of the joint distribution of the second order moments in samples from a bivariate distribution. The general Wishart distribution was stated in (k) of §5.5. We shall now present a derivation of this distribution.

Let O_n: $(x_{11},x_{21},\ldots,x_{k1};\ x_{12},x_{22},\ldots,x_{k2};\ldots;\ x_{1n},x_{2n},\ldots,x_{kn})$ be a sample of n observations from the k-variate population having a p. d. f.

$$\text{(a)} \qquad \frac{\sqrt{A}}{(2\pi)^{k/2}}\, e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}x_i x_j},$$

---

*John Wishart, "The Generalized Product Moment Distribution in Samples from a Normal Multivariate Population", Biometrika, Vol. 20A, pp. 32-52. A proof based on the method of characteristic functions has also been given by J. Wishart and M. S. Bartlett, "The Generalized Product Moment Distribution", Proc. Camb. Phil. Soc., Vol. 29 (1933) pp. 260-270.

where A is the determinant of the positive definite matrix $||A_{ij}||$. Let

$$b_{ij} = \sum_{\alpha=1}^{n} x_{i\alpha}x_{j\alpha}, \qquad (i,j = 1,2,\ldots,k). \tag{b}$$

Clearly $b_{ij} = b_{ji}$, so that there are only $k(k+1)/2$ distinct b_{ij}. The b_{ij}/n may be referred to as second order sample moments. Our problem is to obtain the joint p. d. f. of the b_{ij}. The joint p. d. f. of the $x_{i\alpha}$ $(i = 1,2,\ldots,k;\ \alpha = 1,2,\ldots,n)$ is given by

$$\frac{A^{n/2}}{(2\pi)^{(nk)/2}} e^{-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{i,j=1}^{k} A_{ij}x_{i\alpha}x_{j\alpha}} = \frac{A^{n/2}}{(2\pi)^{(nk)/2}} e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}b_{ij}}, \tag{c}$$

Now, the probability element of the b_{ij} is given by

$$\Pr\left(b_{ij} < \sum_{\alpha=1}^{n} x_{i\alpha}x_{j\alpha} < b_{ij}+db_{ij}; i \leq j = 1,2,\ldots,k\right) = \frac{A^{n/2}}{(2\pi)^{(nk)/2}} \int_R e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}b_{ij}} \prod_{i,\alpha} dx_{i\alpha}, \tag{d}$$

where R is the region in the kn-dimensional space of the $x_{i\alpha}$ for which

$$b_{ij} < \sum_{\alpha=1}^{n} x_{i\alpha}x_{j\alpha} < b_{ij} + db_{ij}, \qquad (i \leq j = 1,2,\ldots,k). \tag{e}$$

within terms of order $\prod_{i\leq j=1}^{k} db_{ij}$, the probability given by (d) may be written as

$$\frac{A^{n/2}}{(2\pi)^{(nk)/2}} e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}b_{ij}} \int_R \prod_{i,\alpha} dx_i \,. \tag{f}$$

Our problem now reduces to the integration of $\prod_{i,\alpha} dx_{i\alpha}$ over the region R. Let $f_1(b_{11})db_{11}$ be the volume element for which $b_{11} < \sum_{1}^{n} x_{1\alpha}^2 < b_{11} + db_{11}$; $f_2(b_{21},b_{22}|b_{11})db_{21}db_{22}$ the volume element for which $b_{21} < \sum_{1}^{n} x_{2\alpha}x_{1\alpha} < b_{21} + db_{21}$, $i = 1,2$, for a fixed value of b_{11}; with a similar meaning for $f_3(b_{31},b_{32},b_{33}|b_{11},b_{21},b_{22})db_{31}db_{32}db_{33}$, and so on. Then the volume element for which $b_{ij} < \sum_{1}^{n} x_{i\alpha}x_{j\alpha} < b_{ij} + db_{ij}$, that is, the integral in (f) (to terms of order $\prod_{i<j} db_{ij}$) is given by the product

$$f_1(b_{11})db_{11}\, f_2(b_{21},b_{22}|b_{11})db_{21}db_{22}\ldots f_k(b_{k1},b_{k2},\ldots,b_{kk}|b_{11},b_{21},b_{22},\ldots b_{k-1\,k-1}) \cdot db_{k1}db_{k2}\ldots db_{kk}. \tag{g}$$

Now, consider the problem of determining the expression for

$$f_m(b_{m1},b_{m2},\ldots,b_{mm}|b_{11},\ldots,b_{m-1,m-1})db_{m1}db_{m2}\ldots db_{mm}. \tag{h}$$

We note that $b_{ij} = \sum_1^n x_{i\alpha}x_{j\alpha}$, $i \leqq j = 1,2,\ldots,m-1$, are fixed. Geometrically, we may represent $P_i(x_{i1},x_{i2},\ldots,x_{in})$, $i = 1,2,\ldots,k$, as k points in an n-dimensional space. $\sqrt{b_{ii}}$ is the distance between the i-th point P_i and the origin O, while $b_{ij}/\sqrt{b_{ii}b_{jj}}$ is the cosine of the angle between the vectors OP_i and OP_j. Fixing b_{ij}, $i \leqq j = 1,2,\ldots,m-1$, means fixing the relative position of the vectors OP_1, $OP_2,\ldots,OP_{m-1}$. The vector OP_m is free to vary in such a way that

$$\text{(i)} \qquad b_{mi} < \sum x_{m\alpha}x_{i\alpha} < b_{mi} + db_{mi}, \qquad (i = 1,2,\ldots,m),$$

and we wish to find the volume of the region over which P_m is free to vary. If $n = m$, we have as many vectors as dimensions and we can find our volume element by making the transformation

$$\sum_1^m x_{m\alpha}x_{i\alpha} = b_{mi}, \qquad (i = 1,2,\ldots,m).$$

The Jacobian is

$$\text{(j)} \qquad \frac{\partial(x_{m1},x_{m2},\ldots,x_{mm})}{\partial(b_{m1},b_{m2},\ldots,b_{mm})} = \frac{1}{\Delta}$$

where

$$\Delta = \begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ 2x_{m1} & 2x_{m2} & \cdots & 2x_{mm} \end{vmatrix} = 2|x_{ij}|, \qquad (i,j = 1,2,\ldots,m).$$

The absolute value of the determinant $|x_{ij}|$ is the volume of the parallelotope based on the edges OP_1, $OP_2,\ldots,OP_m$. By taking the positive square root of the square of this determinant, we may overcome the difficulty of sign. Thus

$$|x_{ij}|^2 = |b_{ij}|$$

and hence

$$\Delta = 2\sqrt{|b_{ij}|}.$$

Therefore, we have

$$\text{(k)} \qquad \prod_{\alpha=1}^m dx_m = \frac{1}{2\sqrt{|b_{ij}|}} \prod_{i=1}^m db_{mi}.$$

Hence the differential element on the right in (k) obtained by taking all values of x_m

for which

$$b_{mi} < \sum_{1}^{m} x_{m\alpha}x_{i\alpha} < b_{mi} + db_{mi}$$

is a function of the volume of the parallelotope and the differentials db_{mi} in the values of the b_{mi}.

It can be shown* that $|b_{ij}|$ is the volume of the parallelotope T_m, based on the edges $OP_1, OP_2, \ldots, OP_m$, for any number of dimensions $n \geq m$. If n exceeds m, then P_m is free to vary within an (n-m+1)-dimensional spherical shell, as will be noted by examining the inequalities in (1). One of these inequalities (i=m) represents an n-dimensional spherical shell of thickness db_{mm}, the remaining inequalities representing pairs of parallel (n-1)-dimensional planes, where in general no two pairs are parallel to each other. The volume included between any arbitrary pair of planes, e. g., $\sum_{1}^{n} x_{m\alpha}x_{i\alpha} = b_{mi}$, and $\sum_{1}^{n} x_{m\alpha}x_{i\alpha} = b_{mi} + db_{mi}$ (i<m) is an m-dimensional slab of thickness $db_{mi}/\sqrt{b_{ii}}$. The intersection of the (m-1) pairs of (n-1)-dimensional planes and the n-dimensional spherical shell yields an (n-m+1)-dimensional spherical shell. Now the inner surface of this shell (or any spherical surface concentric with the inner surface) is perpendicular to the differentials db_{mm}, $db_{m\,m-1}, \ldots, db_{m1}$. This is evident upon examining the manner in which the (n-m+1)-dimensional spherical shell mentioned above is obtained as the common intersection of the m-1 parallel pairs of (n-1)-dimensional planes and the n-dimensional

---

*For example, see D. M. Y. Sommerville An Introduction to the Geometry of n Dimensions, Methuen, London (1929) Chapter 8.

There is also another geometrical interpretation of $|b_{ij}|$ for any $n \geq m$, which is of considerable interest. The $x_{i\alpha}$ ($i = 1,2,\ldots,m$; $\alpha = 1,2,\ldots,n$) may be regarded as n points P_α ($\alpha = 1,2,\ldots,n$) in m dimensions. If we take any m of these n points, say P_{α_r}: $(x_{i\alpha_r}, i = 1,2,\ldots,m)$, $(r = 1,2,\ldots,m)$ together with the origin 0 as the (m+1)-st point, then the square of the volume of the parallelotope based on $OP_{\alpha_1}, OP_{\alpha_2}, \ldots, OP_{\alpha_m}$ is given by $|\sum_{r=1}^{m} x_{i\alpha_r}x_{j\alpha_r}|$. This follows from the discussion between (i) and (j). Now there are ${}_nC_m$ ways of choosing m points from the n points P_α, and hence there are ${}_nC_m$ parallelotopes which can be formed in a manner similar to that discussed above. It can be shown that $|b_{ij}| = \sum' |\sum_{r=1}^{m} x_{i\alpha_r}x_{j\alpha_r}|$, where $\sum'$ denoted summation for all ${}_nC_m$ parallelotopes thus formed. The proof of this follows by mathematical induction by increasing the number of points from m to n successively by unity. In the case i = j = 1, we have n points in one dimension and $|b_{ij}| = b_{11} = \sum_{1}^{n} x_{i\alpha}^2$, the sum of squares of the distances of each point from the origin. In the case of i, j = 1,2,...,k, $|b_{ij}|$ is the sum of squares of the volumes of all ${}_nC_m$ m-dimensional parallelotopes which can be constructed from the n given points, using the origin as one vertex in each parallelotope. $|b_{ij}|$ may be referred to as the generalized sum of squares.

spherical shell. Therefore, the rectangular volume element $db_{m1}db_{m2},\ldots,db_{mm}$ is perpendicular to the inner surface of the (n-m+1)-dimensional spherical shell. The thickness of this shell is given by the differential element $1/(2\sqrt{|b_{ij}|})\prod_1^m db_{mi}$ in (k). Therefore, by multiplying this thickness by the inner surface content of our (n-m+1)-dimensional shell, we obtain the volume of the shell to terms of order $\prod_1^m db_{mi}$. The radius of this inner surface is equal to the distance h from P_m to the (m-1)-dimensional space formed by $OP_1,OP_2,\ldots,OP_{m-1}$. This is, perhaps, seen most readily by noting that the inner surface of our shell is obtained by taking the intersections of $\sum_1^n x_{m\alpha}x_{i\alpha} = b_{mi}$, $i = 1,2,\ldots,m$ (we are assuming, of course, that all db_{mi} are > 0). The center of the sphere having this surface must clearly lie in the intersection of the (m-1) (n-1)-dimensional planes $\sum_1^n x_{m\alpha}x_{i\alpha} = b_{mi}$, $i = 1,2,\ldots,m-1$. This intersection point lies on each of the vectors $OP_1,OP_2,\ldots,OP_{m-1}$, and the line between this point and P_m is perpendicular to each of the first m-1 vectors, which is equivalent to the statement that the center of the (n-m+1)-dimensional shell is at the point where the perpendicular from P_m intersects the (m-1)-dimensional space formed by the remaining m-1 vectors, $OP_1,OP_2,\ldots,OP_{m-1}$. The volume of T_m, the parallelotope formed from $OP_1,OP_2,\ldots,OP_m$, is $\sqrt{|b_{ij}|} = V_m$, say, and that of T_{m-1}, the parallelotope formed from $OP_1,OP_2,\ldots,OP_{m-1}$, is $\sqrt{|b_{\alpha\beta}|} = V_{m-1}$, $\alpha,\beta = 1,2,\ldots,m-1$. Using T_{m-1} as the base of T_m and h as the height, we must have $V_m = hV_{m-1}$, or $h = V_m/V_{m-1}$.

Now the volume of an n-dimensional sphere of radius r is

$$(1)\qquad 2^n \int_0^r \int_0^{\sqrt{r^2-x_1^2}} \cdots \int_0^{\sqrt{r^2-x_1^2-\ldots-x_{n-1}^2}} dx_n dx_{n-1}\cdots dx_1 = \frac{\pi^{\frac{n}{2}} r^n}{\Gamma(\frac{n+2}{2})},$$

and the surface content of the sphere is obtained by taking the derivative of this expression with respect to r, which is found to be

$$(m)\qquad \frac{2\pi^{\frac{n}{2}} r^{n-1}}{\Gamma(\frac{n}{2})}.$$

The integral in (1) may be readily evaluated by integrating immediately with respect to x_n, then setting

$$x_{n-i} = \sqrt{\theta_i}\sqrt{r^2-x_1^2-\ldots-x_{n-i-1}^2},$$

$i = 1,2,\ldots,n-1$, and integrating with respect to the appropriate θ at each stage.

The surface content of the inner surface of our spherical shell is therefore

(n) $$\frac{2\pi^{\frac{n-m+1}{2}}}{\Gamma(\frac{n-m+1}{2})}\left(\frac{V_m}{V_{m-1}}\right)^{n-m},$$

and the content of the spherical shell is obtained by multiplying expression (n) by the thickness $\frac{1}{2V_m}\prod_{i=1}^{m} db_{mi}$. Therefore, we finally obtain as the expression for the function in (h),

(o) $$\frac{\pi^{\frac{n-m+1}{2}}}{\Gamma(\frac{n-m+1}{2})}\frac{V_m^{n-m-1}}{V_{m-1}^{n-m}}\prod_{i=1}^{m} db_{mi}.$$

Letting m take on the values $1,2,\ldots,k$ in (o) and multiplying the results, we obtain the following expression for (g)

(p) $$\frac{\pi^{\frac{kn}{2}-\frac{k(k-1)}{4}}V_k^{n-k-1}}{\prod_{i=1}^{k}\Gamma(\frac{n+1-i}{2})}\prod_{i\leq j} db_{ij},$$

which is the value of $\int_R \prod_{i,\alpha} dx_{i\alpha}$ in (f) to terms of order $\prod_{i\leq j} db_{ij}$. We therefore finally obtain the Wishart distribution:

(q) $$w_{n,k}(b_{ij};A_{ij})\prod_{i\leq j} db_{ij} = \frac{(A/2^k)^{\frac{n}{2}}|b_{ij}|^{\frac{n-k-1}{2}}}{\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}\Gamma(\frac{n+1-i}{2})} e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}b_{ij}}\prod_{i\leq j} db_{ij},$$

which is defined over the region in the b_{ij} space for which $||b_{ij}||$ is positive semi-definite, that is, over all values of the a_{ij} for which $|b_{ij}|$ and all principal minors of all orders are ≥ 0. In order for the distribution to exist it is clear that $n+1-k > 0$.

Since

$$\int w_{n,k}(b_{ij};A_{ij})\prod_{i\leq j} db_{ij} = 1$$

where the integration is taken over the space of the b_{ij}, it is clear that

(r) $$\int |b_{ij}|^{\frac{n-k-1}{2}} e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}b_{ij}}\prod_{i\leq j} db_{ij} = \frac{\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}\Gamma(\frac{n+1-i}{2})}{(A/2^k)^{\frac{n}{2}}}.$$

Replacing A_{ij} by $A_{ij} - 2\theta_{ij}$ ($\theta_{ij} = \theta_{ji}$) in (r) then multiplying the result by

$$\frac{(A/2^k)^{\frac{n}{2}}}{\pi^{\frac{k(k-1)}{4}}\prod_{1}^{k}\Gamma(\frac{n+1-i}{2})}$$

we obtain the m. g. f. of the b_{ii} and $2b_{ij}$ $(i<j)$, which has the value

$$A^{\frac{n}{2}}|A_{ij}-2\theta_{ij}|^{-\frac{n}{2}}. \tag{s}$$

Similarly, the reader may verify that the m. g. f. of the $\sum_1^n x_{i\alpha}^2$ and $2\sum_1^n x_{i\alpha}x_{j\alpha}(i<j)$ as determined from (c) by multiplying (c) by

$$e^{\sum_{i,j}\theta_{ij}(\sum_1^n x_{i\alpha}x_{j\alpha})}$$

and integrating over the entire kn-dimensional space of the x's is also given by (s). Therefore, if one were given the function (q) in advance, one could argue by the multivariate analogue of Theorem (B), §2.81, that it is the distribution function of the $\sum_1^n x_{i\alpha}x_{j\alpha}(=b_{ij})$ where the p. d. f. of the $x_{i\alpha}$ is given by (c).

The Wishart distribution (q) may be regarded as a generalization of the χ^2-distribution to the case of vectors with k-components. In fact for k = 1, the quantity $A_{11}b_{11}$ is distributed according to the χ^2-distribution with n degrees of freedom. In this case b_{11} is the sum of squares of the n sample values of x_1, while in the k-variate case b_{ii} is the sum of squares of the n sample values of the x_i (the i-th component of the vector $x_1,x_2,\ldots,x_k$) and b_{ij} $(i\neq j)$ is the inner product or bilinear form between the n sample values of the x_i and x_j. As in the case of the χ^2-distribution, the Wishart distribution has a reproductive property to be considered in the next section.

11.2 Reproductive Property of the Wishart Distribution.

The reproductive property of Wishart distributions is very useful in multivariate statistical theory, and it may be stated in the following theorem:

Theorem (A). Let $b_{ij}^{(1)},b_{ij}^{(2)},\ldots,b_{ij}^{(p)}$ $(i\leq j = 1,2,\ldots,k)$ be p systems of random variables distributed independently according to Wishart distributions (p. d. f.'s)

$$w_{n_t,k}(b_{ij}^{(t)};A_{ij}), \qquad (t = 1,2,\ldots,p),$$

respectively. Let $b_{ij} = \sum_{t=1}^{p} b_{ij}^{(t)}$, $n = \sum_{t=1}^{p} n_t$. Then the b_{ij} are distributed according to the Wishart p. d. f.

$$w_{n,k}(b_{ij};A_{ij}).$$

To prove this theorem, we determine the m. g. f. $\phi(\theta_{ij})$, $(\theta_{ij}=\theta_{ji})$, of the b_{ii} and $2b_{ij}$ $(i<j)$. We have

(a) $$\phi(\theta_{ij}) = E(e^{\sum_{i,j=1}^{k}\theta_{ij}b_{ij}})$$
$$= E(e^{\sum_{i,j}\theta_{ij}b_{ij}^{(1)} + \sum_{i,j}\theta_{ij}b_{ij}^{(2)} + \cdots + \sum_{i,j}\theta_{ij}b_{ij}^{(p)}})$$
$$= E(e^{\sum_{i,j}\theta_{ij}b_{ij}^{(1)}})\, E(e^{\sum_{i,j}\theta_{ij}b_{ij}^{(2)}}) \cdots E(e^{\sum_{i,j}\theta_{ij}b_{ij}^{(p)}}).$$

But $E(e^{\sum_{i,j}\theta_{ij}b_{ij}^{(t)}}) = A^{\frac{n_t}{2}} |A_{ij}-2\theta_{ij}|^{-\frac{n_t}{2}}$ and therefore

(b) $$\phi(\theta_{ij}) = A^{\frac{n}{2}} |A_{ij}-2\theta_{ij}|^{-\frac{n}{2}},$$

which is the m. g. f. for the Wishart p. d. f.

$$w_{n,k}(b_{ij};A_{ij}),$$

which we conclude, by the multivariate analogue of the Theorem (B), §2.81, to be the distribution of the b_{ij} $(=\sum_{t=1}^{p} b_{ij}^{(t)})$.

11.3 The Independence of Means and Second Order Moments in Samples from a Normal Multivariate Population

Suppose $O_n(x_{i\alpha}, i=1,2,\ldots,k; \alpha=1,2,\ldots,n)$ is a sample from the normal multivariate population having p. d. f.

(a) $$f(x_i) = \frac{\sqrt{A}}{(2\pi)^{k/2}} e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}(x_i-a_i)(x_j-a_j)}.$$

The p. d. f. of the sample is

(b) $$f(x_{i\alpha}) = \frac{A^{\frac{n}{2}}}{(2\pi)^{(kn)/2}} e^{-\frac{1}{2}\sum_{i,j} A_{ij}c_{ij}},$$

where $c_{ij} = \sum_{\alpha=1}^{n}(x_{i\alpha}-a_i)(x_{j\alpha}-a_j)$. Let

(c) $$a_{ij} = \sum_{\alpha=1}^{n}(x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j),$$

where

$$\bar{x}_i = \frac{1}{n}\sum_{\alpha=1}^{n} x_{i\alpha}$$

The a_{ij} are distributed according to the Wishart distribution (q), §11.1, with n replaced by n-1. It was shown in §5.12 for the case k=2 that the $\bar{x}_i$ are distributed according to the normal bivariate law (d), §5.12, and it was remarked that in the general case, the distribution of the $\bar{x}_i$ is given by (e), §5.12. The proof of (e), §5.12, may be carried out by evaluating the m. g. f. of the $(\bar{x}_i - a_i)$, i. e.

$$\text{(d)} \qquad E\left(e^{\sum_1^k \theta_i(\bar{x}_i - a_i)}\right) = \int e^{\sum_1^k \theta_i(\bar{x}_i - a_i)} f(x_{i\alpha}) \prod dx_{i\alpha},$$

where $f(x_{i\alpha})$ is given by (b), the integration being over the entire kn-dimensional space of the $x_{i\alpha}$. The evaluation of this integral may be carried out as an extension of the case of k=2, §5.12. The details are left to the reader. In order to show that the a_{ij} have the Wishart distribution with n replaced by n-1, it is sufficient to show that the m. g. f. of the a_{ii} and $2a_{ij}$ $(i \neq j)$ is $A^{(n-1)/2}|A_{ij} - 2\theta_{ij}|^{-((n-1)/2)}$. The problem of doing this is a direct extension to the k-variate case of the procedure followed for k=2 in §5.5. We shall have to leave the details to the reader.

Just as in the 1 and 2 variable cases discussed in §5.6, the a_{ij} and $\bar{x}_i$ are independently distributed systems. A fairly direct verification of this, although tedious, is to evaluate the joint m. g. f. of the a_{ij} and $\bar{x}_i$ and note that it factors.

11.4 Hotelling's Generalized "Student" Test

Suppose a sample O_n is drawn from a normal multivariate population with distribution (a) in §11.3, and that it is desired to test the hypothesis $H(a_i = a_{io})$ that the a_i have specified values a_{io} $(i=1,2,\ldots,k)$, no matter what values the A_{ij} may have. This hypothesis may be specified as follows:

$$\text{(a)} \qquad \Omega: \begin{cases} A_{ij} \text{ such that } ||A_{ij}|| \text{ is positive definite} \\ \text{and } -\infty < a_i < +\infty,\ i = 1,2,\ldots,k. \end{cases}$$

$$\omega: \begin{cases} \text{The subspace of } \Omega \text{ for which } a_i = a_{io}, \\ \quad i = 1,2,\ldots,k. \end{cases}$$

It will be noted that this is the k-variate analogue of the "Student" statistical hypothesis discussed in §7.2 for one variable, which is simply the hypothesis that a sample from a normal population comes from one having a specified mean, no matter what the variance may be.

The likelihood function for testing the hypothesis $H(a_i = a_{io})$ is given by (b) in §11.3.

Maximizing the likelihood function for variations of the A_{ij} and a_i over Ω, we find

$$\text{(b)} \qquad \hat{a}_i = \bar{x}_i, \quad ||\hat{A}_{ij}|| = ||\frac{a_{ij}}{n}||^{-1},$$

and hence the maximum of the likelihood for variations of the parameters over Ω is

$$\text{(c)} \qquad \frac{1}{(2\pi)^{\frac{nk}{2}} |\frac{a_{ij}}{n}|^{\frac{n}{2}}} e^{-\frac{1}{2}nk}.$$

Similarly, the maximum of the likelihood for variations of the parameters over ω (i. e. for variations of the A_{ij} and for $a_i = a_{io}$) is found to be

$$\text{(d)} \qquad \frac{1}{(2\pi)^{\frac{nk}{2}} |\frac{c_{oij}}{n}|^{\frac{n}{2}}} e^{-\frac{1}{2}nk},$$

where $c_{oij} = c_{ij}$ in (b), §11.3, with $a_i = a_{io}$.

The likelihood ratio for testing $H(a_i = a_{io})$ is the ratio of expression (d) to expression (c), i. e.

$$\text{(e)} \qquad \lambda = \frac{|a_{ij}|^{\frac{n}{2}}}{|c_{oij}|^{\frac{n}{2}}}.$$

Clearly, we may use $\lambda^{\frac{2}{n}} = Y$, say, as a test criterion for $H(a_i = a_{io})$ since it is a single-value function of λ. To complete the derivation of our test, we must determine the distribution of Y when $H(a_i = a_{io})$ is true. We shall obtain this distribution by first finding its moments. Now, we know from §11.1 that the joint distribution of the c_{oij} is the Wishart distribution

$$\text{(f)} \qquad w_{n,k}(c_{oij}; A_{ij}) = \frac{(A/2^k)^{\frac{n}{2}} |c_{oij}|^{\frac{n-k-1}{2}}}{\pi^{\frac{k(k-1)}{4}} \prod_1^k \Gamma(\frac{n+1-i}{2})} \cdot e^{-\frac{1}{2}\sum_{i,j} A_{ij} c_{oij}}.$$

The g-th moment of $|c_{oij}|$ is obtained in the following way. Since the integral of the function (f) over the space S_c of the c_{oij} is unity, we have

$$(g)\qquad \int_{S_c} |c_{oij}|^{\frac{n-k-1}{2}} e^{-\frac{1}{2}\sum_{i,j} A_{ij}c_{oij}} \prod_{i<j} dc_{oij} = \frac{\pi^{\frac{k(k-1)}{4}} \prod_{1}^{k} \Gamma(\frac{n+1-i}{2})}{(A/2^k)^{\frac{n}{2}}}.$$

Replacing n by n+2g in (g), then multiplying by

$$\frac{(A/2^k)^{\frac{n}{2}}}{\pi^{\frac{k(k-1)}{4}} \prod_{1}^{k} \Gamma(\frac{n+1-i}{2})},$$

we obtain an expression on the left which defines $E(|c_{oij}|^g)$ and its value is given on the right. That is

$$(h)\qquad E(|c_{oij}|^g) = \frac{\prod_{1}^{k} \Gamma(\frac{n+1-i}{2} + g)}{(A/2^k)^g \prod_{1}^{k} \Gamma(\frac{n+1-i}{2})}.$$

But the c_{oij} are functions of the a_{ij} and $\bar{x}_i$, since

$$(i)\qquad c_{oij} = a_{ij} + n(\bar{x}_i - a_{io})(\bar{x}_j - a_{jo}).$$

Therefore,

$$(j)\left[\frac{n^{\frac{k}{2}} A^{\frac{1}{2}}}{(2\pi)^{\frac{k}{2}}} \frac{(A/2^k)^{\frac{n-1}{2}}}{\pi^{\frac{k(k-1)}{4}} \prod_{1}^{k} \Gamma(\frac{n-i}{2})}\right] \int_{S_c} |c_{oij}|^g |a_{ij}|^{\frac{n-k-2}{2}} e^{-\frac{1}{2}\sum_{i,j} A_{ij}(a_{ij}+n(\bar{x}_i-a_{io})(\bar{x}_j-a_{jo}))} \prod_{i<j} da_{ij} \cdot \prod d\bar{x}_i$$

$$= \frac{\prod_{1}^{k} \Gamma(\frac{n+1-i}{2} + g)}{(A/2^k)^g \Gamma(\frac{n+1-i}{2})}.$$

Dividing both members of (j) by the expression in [ ], then replacing n by n+2h, except in the distribution of the $\bar{x}_i$ (the n's here being easily removable by changing variables $\sqrt{n}(\bar{x}_i - a_i) = y_i$ say) then multiplying the resulting equation by [ ], we obtain, as the first member, an expression defining $E(|c_{oij}|^g \cdot |a_{ij}|^h)$ and its value is given by the second member; thus

$$(k)\qquad E(|c_{oij}|^g |a_{ij}|^h) = \frac{\prod_{1}^{k} [\Gamma(\frac{n+1-i}{2} + g + h)\Gamma(\frac{n-i}{2} + h)]}{(A/2^k)^{g+h} \prod_{1}^{k} [\Gamma(\frac{n+1-i}{2} + h)\Gamma(\frac{n-i}{2})]}.$$

Clearly this moment will exist for all integers g and h for which all arguments of the

gamma functions are > 0. Setting $g = -h$, we obtain as the h-th moment* of Y,

$$(l) \qquad E(Y^h) = \frac{\prod_1^k [\Gamma(\frac{n+1-i}{2})\Gamma(\frac{n-i}{2} + h)]}{\prod_1^k [\Gamma(\frac{n+1-i}{2} + h)\Gamma(\frac{n-i}{2})]} = \frac{\Gamma(\frac{n}{2})\Gamma(\frac{n-k}{2} + h)}{\Gamma(\frac{n}{2} + h)\Gamma(\frac{n-k}{2})} .$$

This moment may be written as

$$(m) \qquad \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-k}{2})\Gamma(\frac{k}{2})} \frac{\Gamma(\frac{n-k}{2} + h)\Gamma(\frac{k}{2})}{\Gamma(\frac{n}{2} + h)} = \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-k}{2})\Gamma(\frac{k}{2})} \int_0^1 x^{\frac{n-k}{2} + h - 1} (1-x)^{\frac{k}{2} - 1} dx.$$

Therefore the h-th moment of Y ($h = 0,1,2,\ldots$) is identical with the h-th moment of a variable x having probability element

$$(n) \qquad \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-k}{2})\Gamma(\frac{k}{2})} x^{\frac{n-k}{2} - 1} (1-x)^{\frac{k}{2} - 1} dx.$$

It follows from Theorem (A), §2.76, on the uniqueness of distributions from moments that Y is distributed according to the probability law (n).

Making use of the fact that

$$c_{oij} = a_{ij} + \sqrt{n}(\bar{x}_i - a_{io}) \cdot \sqrt{n}(\bar{x}_j - a_{jo})$$

and letting

$$y_i = \sqrt{n}(\bar{x}_i - a_{io})$$

we may write

$$(o) \qquad |c_{oij}| = |a_{ij} + y_i y_j|$$

$$\begin{vmatrix} -1 & y_1 & y_2 & \cdots & y_k \\ 0 & a_{11}+y_1^2 & a_{12}+y_1y_2 & \cdots & a_{1k}+y_1y_k \\ 0 & a_{21}+y_2y_1 & a_{22}+y_2^2 & \cdots & a_{2k}+y_2y_k \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & a_{k1}+y_ky_1 & a_{k2}+y_ky_2 & \cdots & a_{kk}+y_k^2 \end{vmatrix}$$

*For more applications of the foregoing technique of finding moments of ratios of determinants, see S. S. Wilks "Certain Generalizations in the Analysis of Variance", Biometrika, Vol. 24 (1932) pp. 471-494.

Multiplying the first row by $-y_1$ then adding to the second; multiplying the first row by $-y_2$ and adding to the third; and so on, we may write the determinant as

$$(p) \qquad - \begin{vmatrix} -1 & y_1 & y_2 & \cdots & y_k \\ y_1 & a_{11} & a_{12} & \cdots & a_{1k} \\ y_2 & a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ y_k & a_{k1} & a_{k2} & \cdots & a_{kk} \end{vmatrix} .$$

It follows from the argument leading to expression (k), §3.23 that the expression (p) may be written as $a_{ij}[1+\sum_{i,j=1}^{k} a^{ij}y_iy_j]$, and substituting the value of y_i we are finally able to write

$$(q) \qquad |c_{oij}| = |a_{ij}|[1+n\sum_{i,j=1}^{k} a^{ij}(\bar{x}_i-a_{io})(\bar{x}_j-a_{jo})] = |a_{ij}|[1+\frac{T^2}{n-1}],$$

where T^2 is Hotelling's* Generalized Student Ratio which can be written down explicitly in terms of the a^{ij} and $(\bar{x}_i-a_{io})$ in an obvious way. Hence

$$(r) \qquad Y = \frac{1}{1+T^2/n-1} ,$$

and the distribution of T can be found at once by applying the transformation (r) to the probability element (n) (with x replaced by Y). The result is

$$(s) \qquad \frac{2\Gamma(\frac{n}{2})}{\Gamma(\frac{n-k}{2})\Gamma(\frac{k}{2})\sqrt{n-1}} \frac{(T^2/n-1)^{\frac{k-1}{2}}\,dT}{(1+T^2/n-1)^{n/2}} .$$

11.5 The Hypothesis of Equality of Means in Multivariate Normal Populations

Suppose O_{n_t} $(x^t_{i\alpha}$, $i=1,2,\ldots,k$; $\alpha=1,2,\ldots,n_t$; $t=1,2,\ldots,p)$ are p samples from the normal k-variate populations

$$(a) \qquad \frac{A^{\frac{n_t}{2}}}{(2\pi)^{\frac{n_tk}{2}}} e^{-\frac{1}{2}\sum_{i,j} A_{ij}(x^t_i-a^t_i)(x^t_j-a^t_j)} , \qquad (t=1,2,\ldots,p),$$

and that it is desired to test the following hypothesis:

---

*H. Hotelling, "The Generalization of Student's Ratio", Annals of Math. Stat., Vol. 2 (1931) pp. 359-378.

(b)
$$\Omega:\begin{cases} A_{ij} \text{ such that } ||A_{ij}|| \text{ is positive definite} \\ \text{and } -\infty < a_i^t < \infty,\ i=1,2,\ldots,k;\ t=1,2,\ldots,p. \end{cases}$$
$$\omega:\begin{cases} \text{Subspace of } \Omega \text{ for which } a_i^1 = a_i^2 = \ldots = a_i^p = a_i, \\ \text{where } -\infty < a_i < \infty,\ i = 1,2,\ldots,k. \end{cases}$$

Denoting this hypothesis by $H(a_i^1=a_i^2=\ldots=a_i^p)$, it is simply the hypothesis that the samples come from k-variate normal populations having identical sets of means, given that they come from k-variate normal populations with the same variance-covariance matrix. It should be noted that this hypothesis is the multivariate analogue of that treated in §9.1.

Let

(c)
$$a_{ij}^t = \sum_{\alpha=1}^{n_t}(x_{i\alpha}^t - \bar{x}_i^t)(x_{j\alpha}^t - \bar{x}_j^t),$$
$$\bar{a}_{ij} = \sum_{t=1}^{p} a_{ij}^t,$$

and

(d)
$$a_{ij} = \sum_{t=1}^{p}\sum_{\alpha=1}^{n_t}(x_{i\alpha}^t - \bar{x}_i)(x_{j\alpha}^t - \bar{x}_j),$$

where

(e)
$$\bar{x}_i^t = \frac{1}{n_t}\sum_{\alpha=1}^{n_t} x_{i\alpha}^t,$$

and

(f)
$$\bar{x}_i = \frac{1}{n}\sum_{t=1}^{p}\sum_{\alpha=1}^{n_t} x_{i\alpha}^t = \frac{1}{n}\sum_{t=1}^{p} n_t \bar{x}_i^t.$$

The a_{ij}/n are the second-order product moments in the pool of all samples, and similarly $\bar{x}_i$ is the mean of the i-th variate in the pool of all samples.

The likelihood function for all samples is

(g)
$$\frac{A^{\frac{n}{2}}}{(2\pi)^{\frac{nk}{2}}}\, e^{-\frac{1}{2}\sum_{i,j} A_{ij} c_{ij}^t},$$

where

$$c_{ij}^t = \sum_{\alpha=1}^{n_t}(x_{i\alpha}^t - a_i^t)(x_{j\alpha}^t - a_j^t).$$

Maximizing the likelihood function for variations of all parameters over Ω, we obtain

$$\hat{a}_i^t = \bar{x}_i^t,\quad ||\hat{A}_{ij}|| = \left|\left|\frac{\bar{a}_{ij}}{n}\right|\right|^{-1}$$

and the maximum of the likelihood turns out to be

$$\text{(h)} \qquad \frac{1}{(2\pi)^{\frac{nk}{2}} \left|\frac{\bar{a}_{ij}}{n}\right|^{\frac{n}{2}}} e^{-\frac{nk}{2}}.$$

Similarly, maximizing the likelihood function for variations of the parameters over ω, we obtain

$$\text{(i)} \qquad \hat{a}_i = \bar{x}_i, \quad ||\hat{A}_{ij}|| = \left|\left|\frac{a_{ij}}{n}\right|\right|^{-1},$$

and the maximum of the function turns out to be

$$\text{(j)} \qquad \frac{1}{(2\pi)^{\frac{nk}{2}} \left|\frac{a_{ij}}{n}\right|^{\frac{n}{2}}} e^{-\frac{nk}{2}}.$$

Hence the likelihood ratio for testing $H(a_1^1=a_1^2=\ldots=a_1^p)$ is the ratio of (j) to (h), i. e.

$$\text{(k)} \qquad \lambda = \frac{|\bar{a}_{ij}|^{\frac{n}{2}}}{|a_{ij}|^{\frac{n}{2}}}.$$

Again we may use $\lambda^{2/n} = Z$, say, as our test criterion. To find the distribution of Z, we proceed as in §11.4 by the method of moments. Noting that the a_{ij} are distributed according to the Wishart distribution $w_{n-1,k}(a_{ij};A_{ij})$ we have, similar to (h), §11.4,

$$\text{(l)} \qquad E(|a_{ij}|^g) = \frac{\prod_1^k \Gamma(\frac{n-i}{2} + g)}{(A/2^k)^g \prod_1^k \Gamma(\frac{n-i}{2})}.$$

Now it may be verified that

$$\text{(m)} \qquad a_{ij} = \bar{a}_{ij} + m_{ij},$$

where $m_{ij} = \sum_{t=1}^{p} \sum_{i,j=1}^{k} n_t(\bar{x}_i^t-\bar{x}_1)(\bar{x}_j^t-\bar{x}_j)$. Since the $\bar{a}_{ij}$ are functions only of the a_{ij}^t, the a_{ij}^t and $\bar{x}_i^t$ being independently distributed systems, it follows that the $\bar{a}_{ij}$ and the $\bar{x}_i^t$ are independently distributed systems. The a_{ij}^t are distributed according to Wishart distributions $w_{n_t-1,k}(a_{ij}^t;A_{ij})$, t=1,2,...,p, and it follows from the reproductive property of the Wishart distribution that the $\bar{a}_{ij}$ are distributed according to $w_{n-p,k}(\bar{a}_{ij},A_{ij})$. Therefore by using the joint distribution of the $\bar{a}_{ij}$ and $\bar{x}_i^t$ and following steps similar to those yielding (k) in §11.4, we find

(n) $$E(|a_{ij}|^g|\bar{a}_{ij}|^h) = \frac{\prod_1^k [\Gamma(\frac{n-1}{2}+g+h)\Gamma(\frac{n-p+1-1}{2}+h)]}{(A/2^k)^{g+h}\prod_1^k [\Gamma(\frac{n-1}{2}+h)\Gamma(\frac{n-p+1-1}{2})]}.$$

The h-th moment of Z is given by setting g = -h. We find

(o) $$E(Z^h) = \frac{\prod_1^k [\Gamma(\frac{n-1}{2})\Gamma(\frac{n-p+1-1}{2}+h)]}{\prod_1^k [\Gamma(\frac{n-1}{2}+h)\Gamma(\frac{n-p+1-1}{2})]}.$$

It should be noted that for the case of two samples, (i. e. p=2), the h-th moment of Z reduces to

(p) $$\frac{\Gamma(\frac{n-1}{2})\Gamma(\frac{n-k-1}{2}+h)}{\Gamma(\frac{n-1}{2}+h)\Gamma(\frac{n-k-1}{2})},$$

and hence the distribution of Z in this case is the same as that of Y with n replaced by n-1. In the two-sample case, it should be remembered that $n = n_1 + n_2$, the sum of the two sample numbers.

For the case of p = 3, the h-th moment of Z is

(q) $$\frac{\Gamma(\frac{n-1}{2})\Gamma(\frac{n-2}{2})\Gamma(\frac{n-k-1}{2}+h)\Gamma(\frac{n-k-2}{2}+h)}{\Gamma(\frac{n-1}{2}+h)\Gamma(\frac{n-2}{2}+h)\Gamma(\frac{n-k-1}{2})\Gamma(\frac{n-k-2}{2})} \cdots$$

Making use of the formula

(r) $$\Gamma(i+\tfrac{1}{2})\Gamma(i+1) = \frac{\sqrt{\pi}\,\Gamma(2i+1)}{2^{2i}},$$

(q) reduces to

(s) $$\frac{\Gamma(n-2)\Gamma(n-k-2+2h)}{\Gamma(n-2+2h)\Gamma(n-k-2)}$$

from which we infer the distribution of Z to be identical with that of x^2, where x is distributed according to

(t) $$\frac{\Gamma(n-2)}{\Gamma(n-k-2)\Gamma(k)} x^{n-k-3}(1-x)^{k-1}dx.$$

Setting $Z = x^2$, we find $dx = \frac{1}{2}Z^{-(1/2)}dZ$, and hence the distribution of Z for the case of three samples is

(u) $$\frac{\Gamma(n-2)}{2\Gamma(n-k-2)\Gamma(k)} Z^{\frac{n-k-4}{2}}(1-\sqrt{Z})^{k-1}dZ.$$

The distributions for $p = 4$ and 5 turn out to be relatively simple also.

11.6 The Hypothesis of Independence of Sets of Variables in a Normal Multivariate Population

Suppose $O_n(x_{i\alpha}, i=1,2,\ldots,k; \alpha=1,2,\ldots,n)$ is a sample from a normal multivariate population with distribution (a) in §11.3. Let the variates x_i be grouped into r groups as follows: G_1: $(x_1,x_2,\ldots,x_{k_1})$, G_2: $(x_{k_1+1},x_{k_1+2},\ldots,x_{k_1+k_2})$,..., G_r: $(x_{k_1+\ldots+k_{r-1}+1},\ldots,x_k)$, where $k = k_1+k_2+\ldots+k_r$. The problem we wish to consider is that of deriving a test for the hypothesis that these groups of variates are mutually independent, i. e. that $A_{ij} = 0$ for all i,j not belonging to the same group of variates. Let $||A_{ij}^{(o)}||$ denote the value of $||A_{ij}||$ when all A_{ij} are 0 for i,j belonging to different groups of variates. The hypothesis to be tested may then be specified as follows:

(a)
$$\Omega: \begin{cases} \text{Space of the } A_{ij} \text{ such that } ||A_{ij}|| \text{ is positive} \\ \text{definite and } -\infty < a_i < +\infty. \end{cases}$$
$$\omega: \quad \text{Subspace of } \Omega \text{ for which } ||A_{ij}|| = ||A_{ij}^{(o)}||.$$

We denote this hypothesis by $H(||A_{ij}|| = ||A_{ij}^{(o)}||)$. Maximizing the likelihood function (b) in §11.2 for variations of the parameters over Ω, we find the maximum to be

(b)
$$\frac{1}{(2\pi)^{\frac{nk}{2}} \left|\frac{a_{ij}}{n}\right|^{\frac{n}{2}}} e^{-\frac{1}{2}nk}.$$

The maximum of the likelihood function for variations of the parameters under ω is

(c)
$$\frac{1}{(2\pi)^{\frac{nk}{2}} \left|\frac{a_{ij}^{(o)}}{n}\right|^{\frac{n}{2}}} e^{-\frac{1}{2}nk},$$

where $a_{ij}^{(o)} = a_{ij}$ if i,j belong to the same group of variates and $a_{ij}^{(o)} = 0$ if i,j belong to different groups of variates. Clearly $|a_{ij}^{(o)}|$ is equal to the product of r mutually exclusive principal minors $\prod_{u=1}^{r} |a_{i_u j_u}|$, the u-th minor being the determinant of all a_{ij} associated with the u-th group of variates. Similarly $|A_{ij}^{(o)}| = \prod_{u=1}^{r} |A_{i_u j_u}|$. The likelihood ratio for testing $H(||A_{ij}|| = ||A_{ij}^{(o)}||)$ is, therefore,

(d)
$$\lambda = \frac{|a_{ij}|^{\frac{n}{2}}}{|a_{ij}^{(o)}|^{\frac{n}{2}}}$$

Denoting $\Lambda^{\frac{2}{n}}$ by W, which may clearly be used as the test criterion in place of Λ, we determine the distribution of W by the method of moments.

It should be noted that if we factor $\sqrt{a_{ii}}$ out of the i-th row and i-th column, (i=1,2,...,k) of each of the two determinants $|a_{ij}|$ and $|a_{ij}^{(o)}|$, and using the fact that $a_{ij}/\sqrt{a_{ii}a_{jj}} = r_{ij}$, the sample correlation coefficient between the i-th and j-th variates, then

$$W = \frac{|r_{ij}|}{|r_{ij}^{(o)}|},$$

where $r_{ii} = 1$, and $r_{ij}^{(o)} = r_{ij}$ if i and j both belong to the same group of variates, and $r_{ij}^{(o)} = 0$ if i and j belong to different groups of variates.

To find the moments of W, let us divide the a_{ij} into two classes: (A) those for which i and j correspond to different groups of variates, and (B) all others. Let the product of differentials of the a_{ij} in (A) be dV_A with a similar meaning for dV_B. Now it is evident that if we integrate the Wishart distribution $w_{n,k}(a_{ij};A_{ij}^{(o)})$ with respect to all a_{ij} in Class (A), we will obtain the product of Wishart distributions

$$\prod_{u=1}^{r} w_{n,k_u}(a_{i_u j_u};A_{i_u j_u}),$$

since this integration simply yields the joint distribution of the a_{ij} in Class (B) which we know to be independently distributed in sets $a_{i_u j_u}$ (u=1,2,...,r) when $||A_{ij}|| = ||A_{ij}^{(o)}||$, each set being distributed according to a Wishart law. Hence we must have

$$\text{(e)}\qquad \frac{|A_{ij}^{(o)}|^{\frac{n-1}{2}}}{2^{\frac{(n-1)k}{2}}\pi^{\frac{k(k-1)}{4}}\prod_{1}^{k}\Gamma(\frac{n-1}{2})}\int |a_{ij}|^{\frac{n-k-2}{2}} e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij}^{(o)} a_{ij}} dV_A$$

$$= \prod_{u=1}^{r}\left[\frac{|A_{i_u j_u}|^{\frac{n-1}{2}}}{2^{\frac{(n-1)k_u}{2}}\pi^{\frac{k_u(k_u-1)}{4}}\prod_{i=1}^{k}\Gamma(\frac{n-i}{2})} |a_{i_u j_u}|^{\frac{n-k_u-2}{2}} e^{-\frac{1}{2}\sum_{i_u j_u} A_{i_u j_u} a_{i_u j_u}}\right].$$

Let both members of (e) be multiplied by $\prod_{u=1}^{r}|a_{i_u j_u}|^{-h}$ (which is constant as far as the a_{ij} in Class (A) are concerned), then replace n by n + 2h throughout (e), then multiply throughout by

$$\text{(f)}\qquad \frac{\prod_{1}^{k}\Gamma(\frac{n-i}{2}+h)\, 2^{hk}}{|A_{ij}^{(o)}|^{h}\prod_{1}^{k}\Gamma(\frac{n-i}{2})}$$

then integrate with respect to the a_{ij} in (B). It will be seen that the first member in (e) after these operations will be the integral expression defining $E(W^h)$, and the second member will be the value of $E(W^h)$. We find

$$(g) \qquad E(W^h) = \prod_{u=1}^{r} \prod_{i=1}^{k_u} \left[\frac{\Gamma(\frac{n-i}{2})}{\Gamma(\frac{n-i}{2}+h)} \right] \cdot \prod_{i=1}^{k} \left[\frac{\Gamma(\frac{n-i}{2}+h)}{\Gamma(\frac{n-i}{2})} \right].$$

As a special case, suppose we wish to test the hypothesis that x_1 is independent of the set $x_2, x_3, \ldots, x_k$. In this case $r = 2$, $k_1 = 1$, $k_2 = k-1$. The W criterion is

$$(h) \qquad W = \frac{|r_{ij}|}{\bar{r}_{11}} = 1 - R^2,$$

where $\bar{r}_{11}$ is the minor of the element in the first row and column of $|r_{ij}|$, and R is the sample multiple correlation coefficient between x_1 and $x_2, x_3, \ldots, x_k$. The h-th moment of W for this case is found from (g) to be

$$(i) \qquad \frac{\Gamma(\frac{n-1}{2})\Gamma(\frac{n-k}{2}+h)}{\Gamma(\frac{n-1}{2}+h)\Gamma(\frac{n-k}{2})}.$$

Following the procedure used in inferring the distribution of Y in §11.4 from its h-th moment, we find the probability element of W to be

$$(j) \qquad \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-k}{2})\Gamma(\frac{k-1}{2})} W^{\frac{n-k-2}{2}} (1-W)^{\frac{k-3}{2}} dW.$$

Setting $W = 1-R^2$, we easily find the distribution law of R^2, the square of the sample multiple correlation coefficient, between x_1 and $x_2, x_3, \ldots, x_k$ to be

$$(k) \qquad \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-k}{2})\Gamma(\frac{k-1}{2})} (R^2)^{\frac{k-3}{2}} (1-R^2)^{\frac{n-k-2}{2}} d(R^2),$$

when the hypothesis of independence of x_1 and $x_2, x_3, \ldots, x_k$ is true, i. e. when the $A_{1j}=0$, $(j=2,3,\ldots,k)$, which is equivalent to having the multiple correlation coefficient equal to zero in the population. This result was first obtained by R. A. Fisher, who also

later* derived the distribution of R^2 in samples from a normal multivariate population having an arbitrary multiple correlation coefficient.

Distributions of W for various special cases involving two and three groups of variates have been given by Wilks**.

11.7 Linear Regression Theory in Normal Multivariate Populations

The theorems and other results presented in Chapters VIII and IX can be extended to the case in which the dependent variable y is a vector with an arbitrary number of components (say $y_1, y_2, \ldots, y_s$), each component being distributed normally about a linear function of the fixed variates $x_1, x_2, \ldots, x_k$. In this section we shall state without proof the multivariate analogues of the important theorems in Chapter VIII. The details of the proofs*** of these theorems are rather tedious and can be carried out as extensions of the proofs for the case of one variable.

Suppose $y_1, y_2, \ldots, y_k$ are distributed according to the normal multivariate distribution

$$\text{(a)} \qquad \frac{A^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}\sum_{i,j=1}^{s} A_{ij}(y_i-b_i)(y_j-b_j)},$$

where

$$\text{(b)} \qquad b_i = \sum_{p=1}^{k} a_{ip}x_p,$$

the x_p being fixed variates. Let $O_n\,(y_{i\alpha}|x_{p\alpha};\ i=1,2,\ldots,s;\ p=1,2,\ldots,k;\ \alpha=1,2,\ldots,n)$ be

---

*R. A. Fisher, "The General Sampling Distribution of the Multiple Correlation Coefficient" Proc. Roy. Soc. London, Vol. 121 (1928) pp. 654-673.
An alternative derivation has also been given by S. S. Wilks, "On the Sampling Distribution of the Multiple Correlation Coefficient", Annals Math. Stat., Vol. 3 (1932) pp. 196-203.

**S. S. Wilks, "On the Independence of k Sets of Normally Distributed Statistical Variables", Econometrica, Vol. 3 (1935), pp. 309-326.

***Proofs and extensions of many of the results may be found in one or more of the following papers:
M. S. Bartlett, "On the Theory of Statistical Regression", Proc. Royal Soc. Edinburgh, Vol. 53 (1933), pp. 260-283.
P. L. Hsu, "On Generalized Analysis of Variance", Biometrika, Vol. 31 (1940), pp. 221-237.
D. N. Lawley, "A Generalization of Fisher's z", Biometrika, Vol. 30 (1938), pp. 180-187.
W. G. Madow, "Contributions to The Theory of Multivariate Statistical Analysis", Trans. Amer. Math. Soc., Vol. 44 (1938), pp. 454-495.
S. S. Wilks, "Moment-Generating Operators for Determinants of Product Moments in Samples from a Normal System", Annals of Math., Vol. 35 (1934), pp. 312-340.

a sample from a population having distribution (a). The likelihood function associated with this sample is

$$\text{(c)} \qquad \frac{A^{\frac{n}{2}}}{(2\pi)^{\frac{ns}{2}}} e^{-\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{i,j=1}^{s} A_{ij}(y_{i\alpha}-b_{i\alpha})(y_{j\alpha}-b_{j\alpha})},$$

where

$$\text{(d)} \qquad b_{i\alpha} = \sum_{p=1}^{k} a_{ip}x_{p\alpha}.$$

Let

$$c_{ij} = \sum_{\alpha=1}^{n} y_{i\alpha}y_{j\alpha},$$

$$\text{(e)} \qquad c'_{iq} = \sum_{\alpha=1}^{n} y_{i\alpha}x_{q\alpha}, \quad c''_{pq} = \sum_{\alpha=1}^{n} x_{p\alpha}x_{q\alpha}.$$

Clearly $c_{ij} = c_{ji}$, $c'_{iq} = c'_{qi}$, $c''_{pq} = c''_{qp}$. For a given value of i, let $\hat{a}_{ip}$ be the solution of the equations

$$\text{(f)} \qquad c'_{iq} - \sum_{p=1}^{k} a_{ip}c''_{pq} = 0, \qquad (q=1,2,\ldots,k),$$

that is,

$$\text{(g)} \qquad \hat{a}_{ip} = \sum_{q=1}^{k} c''^{pq}c'_{iq},$$

and let

$$\text{(h)} \qquad \hat{b}_{i\alpha} = \sum_{p=1}^{k} \hat{a}_{ip}x_{p\alpha}.$$

Furthermore, let

$$\text{(i)} \qquad s_{ij} = \sum_{\alpha=1}^{n} (y_{i\alpha}-\hat{b}_{i\alpha})(y_{j\alpha}-\hat{b}_{j\alpha}).$$

The essential functional and probability properties of the quantities defined in (d), (e), (f), (g), (h) and (i) may be stated in the following theorems:

Theorem (A):

$$\text{(j)} \quad \sum_{\alpha=1}^{n}\sum_{i,j=1}^{s} A_{ij}(y_{i\alpha}-b_{i\alpha})(y_{j\alpha}-b_{j\alpha}) = \sum_{i,j=1}^{s} A_{ij}s_{ij} + \sum_{p,q=1}^{k}\sum_{i,j=1}^{s} A_{ij}c''_{pq}(\hat{a}_{ip}-a_{ip})(\hat{a}_{jq}-a_{jq}).$$

Theorem (B):

$$|s_{ij}| = \begin{vmatrix} c_{ij} & \vdots & c'_{iq} \\ \cdots & \cdots & \cdots \\ c'_{pj} & \vdots & c''_{pq} \end{vmatrix} \Big/ |c''_{pq}| \quad (i,j=1,2,\ldots,s;\ p,q=1,2,\ldots,k).$$

Theorem (C): If O_n: $(y_{i\alpha}|x_{p\alpha})$ is a sample from a population having distribution (a), then if the $x_{p\alpha}$ are such that $||c''_{pq}||$ is positive definite, the s_{ij} are distributed according to the Wishart distribution

$$(k) \qquad w_{n-k,s}(s_{ij};A_{ij}),$$

and independently of the $\hat{a}_{ip}$ $(i=1,2,\ldots,s;\ p=1,2,\ldots,k)$ which are distributed according to the normal ks-variate distribution law

$$(l) \qquad \frac{\sqrt{D}}{\pi^{\frac{ks}{2}}}\, e^{-\frac{1}{2}\sum_{p,q=1}^{k}\sum_{i,j=1}^{s} A_{ij}c''_{pq}(\hat{a}_{ip}-a_{ip})(\hat{a}_{jq}-a_{jq})},$$

where D is the ks-order determinant $|A_{ij}c''_{pq}|$ and has the value $A^k\cdot|c''_{pq}|^s$.

The multivariate analogue of the general linear regression hypothesis stated in §8.3 may be specified as follows:

$$(m) \qquad \begin{cases} \Omega\text{: The space for which } \|A_{ij}\| \text{ is positive definite and } -\infty < a_{ip} < \infty,\ i=1,2,\ldots,s;\ p=1,2,\ldots,k. \\ \omega\text{: The subspace of } \Omega \text{ for which } a_{ip} = a_{ipo},\ i=1,2,\ldots,s;\ p=r+1,\ldots,k. \end{cases}$$

Let us denote this hypothesis by $H(a_{ip}=a_{ipo})$. It is the hypothesis that the last k-r regression coefficients corresponding to y_i $(i=1,2,\ldots,s)$ have specified values a_{ipo}. If the $a_{ipo} = 0$, our hypothesis is that each y_i is independent of $x_{r+1},x_{r+2},\ldots,x_k$.

The likelihood ratio λ for testing this hypothesis (as obtained by maximizing the likelihood (c) for variations of the parameters over Ω and by maximizing for variations of the parameters over ω and taking the ratio of the two maxima) turns out to be given by $U^{n/2}$, where

$$(n) \qquad U = \frac{|s_{ij}|}{|s^*_{ij}|}.$$

The form of s^*_{ij} may be seen from the following considerations:

In view of Theorem (A), when the likelihood function (c) is maximized for variations of the parameters over ω, we may consider the maximizing process in two steps: First, with respect to the a_{ip} parameters over ω (holding the A_{ij} fixed). Here we fix $a_{ip} = a_{ipo}$, $(i=1,2,\ldots,s;\ p=r+1,\ldots,k)$ in (j) and minimize the second term on the right side of (j) with respect to a_{ip} $(i=1,2,\ldots,s;\ p=1,2,\ldots,r)$. The coefficient of A_{ij} in the

right hand side of (j) after this minimizing step is s^*_{ij}, where

(o) $$s^*_{ij} = s_{ij} + m_{ij},$$

where m_{ij} results from the second term of the right hand side of (j). We next maximize (o) for variations of the A_{ij} after maximizing with respect to the a_{ip} (over ω). It will be seen that the maximizing values $\hat{A}_{ij}$ of A_{ij} are obtainable after the first maximizing step (i. e. with respect to the a_{ip} over ω), and are given by

$$||A_{ij}|| = ||\frac{s^*_{ij}}{n}||^{-1}.$$

It will be noted that the form of s^*_{ij} is similar to that of s_{ij}, and is given by

(p) $$s^*_{ij} = \sum_{\alpha=1}^{n}(\dot{y}_{i\alpha}-\hat{b}^*_{i\alpha})(\dot{y}_{j\alpha}-\hat{b}^*_{j\alpha}),$$

where

$$\dot{y}_{i\alpha} = y_{i\alpha} - \sum_{p=r+1}^{k} a_{ipo}x_{p\alpha}, \qquad \hat{b}^*_{i\alpha} = \sum_{p=1}^{r}\hat{a}^*_{ip}x_{p\alpha},$$

and where $\hat{a}^*_{ip}$ are given solving the equations

(q) $$c'^*_{iq} - \sum_{p=1}^{r} a^*_{ip}c''_{pq} = 0, \qquad (q=1,2,\ldots,r),$$

where

(r) $$c'^*_{iq} = \sum_{\alpha=1}^{n}\dot{y}_{i\alpha}x_{q\alpha}.$$

The m_{ij} in (o) are functions of the $\hat{a}_{ip}$ which are distributed independently of the s_{ij}. In fact, it can be shown that the m_{ij} are of the form $\sum_{u=1}^{k-r}\xi_{iu}\xi_{ju}$, where the ξ_{iu} $(i=1,2,\ldots,s)$ are linear functions of the $\hat{a}_{ip}$ distributed according to

(s) $$\frac{\sqrt{A}}{(2\pi)^{\frac{k}{2}}}\, e^{-\frac{1}{2}\sum_{i,j=1}^{s} A_{ij}\xi_{iu}\xi_{ju}}$$

and furthermore the sets ξ_{iu} $(u=1,2,\ldots,k-r)$ are independently distributed, and are distributed independently of the s_{ij} where $H(a_{ip}=a_{ipo})$ is true. If the $a_{ipo} = 0$ $(i=1,2,..,s$ $p=r+1,\ldots,k)$ then it follows from Theorem (B) that $|s^*_{ij}|$ may be expressed as the ratio of two determinants as follows:

(t) $$|s^{*}_{ij}| = \begin{vmatrix} c_{ij} & c'_{iq'} \\ c'_{p'j} & c''_{p'q'} \end{vmatrix} \Big/ |c''_{p'q'}|. \qquad \begin{matrix} (i,j=1,2,\ldots,s; \\ p',q'=1,2,\ldots,r). \end{matrix}$$

Now the problem of determining the distribution of U when $H(a_{ip}=a_{ipo})$ is true, is, therefore, reduced to that of determining the distribution of the ratio of determinants

(u) $$\frac{|s_{ij}|}{|s_{ij} + \sum_{u=1}^{k-r} \xi_{iu}\xi_{ju}|},$$

where the s_{ij} are distributed according to the Wishart distribution

(v) $$w_{n-k,s}(s_{ij};A_{ij}),$$

and the ξ_{iu} are distributed according to

(w) $$\frac{A^{\frac{k-r}{2}}}{(2\pi)^{\frac{(k-r)s}{2}}} e^{-\frac{1}{2}\sum_{u=1}^{k-r}\sum_{i,j=1}^{s} A_{ij}\xi_{iu}\xi_{ju}},$$

the s_{ij} and ξ_{iu} being independently distributed systems.

The simplest procedure for finding the distribution of U is perhaps by the method of moments. The method of finding the moments of U is entirely similar to that of finding the moments of Y and Z in §11.4 and §11.5, respectively. The h-th moment is given by

(x) $$E(U^h) = \frac{\prod_1^s [\Gamma(\frac{n-r+1-i}{2})\Gamma(\frac{n-k+1-i}{2}+h)]}{\prod_1^s [\Gamma(\frac{n-r+1-i}{2}+h)\Gamma(\frac{n-k+1-i}{2})]},$$

from which one may infer the distribution of U in any given case by methods illustrated in §11.5. We may summarize our remarks in the following theorem which is the multivariate analogue of Theorem (A), §8.3.

Theorem (D): Let $O_n(y_{i\alpha}|x_{p\alpha})$ be a sample of size n from the population having distribution (c). Let $H(a_{ip}=a_{ipo})$ be the statistical hypothesis specified by (m), and let $U = \lambda^{2/n}$, where λ is the likelihood ratio for testing the hypothesis. Then

(y) $$U = \frac{|s_{ij}|}{|s_{ij}+m_{ij}|}, \qquad (i,j=1,2,\ldots,s),$$

where s_{ij} is defined by (i), and m_{ij} by (o) and (p), and if $H(a_{ip}=a_{ipo})$ is true, the h-th moment of U is given by (x).

It should be observed that U is a generalized form of the ratio

$$\frac{n\sigma_\Omega^2}{n\sigma_\Omega^2 + n(\sigma_\omega^2-\sigma_\Omega^2)}$$

in Theorem (A), §8.3. In fact, when $s = 1$, then $s_{11} = n\sigma_\Omega^2$ and $m_{11} = n(\sigma_\omega^2-\sigma_\Omega^2)$.

It may be verified that Theorem (D) is general enough to cover multivariate analogues of Case 1 (§8.41), Case 2 (§8.42) and Case 3 (§8.43). The essential point to be noted in all of these cases is that k represents the number of functionally independent a_{ip} (for each i) involved in specifying Ω and r ($<k$) represents the number of functionally independent a_{ip} (for each i) involved in specifying ω.

11.8 Remarks on Multivariate Analysis of Variance Theory

The application of normal linear regression theory to the various analysis of variance problems discussed in Chapter IX can be extended in every instance to the case in which the dependent variable y is a vector of several components. In all such multivariate extensions, U in §11.7 plays the role in making significance tests analogous to Snedecor's F (or $1/(1+\frac{k-r}{n-k}F)$, to be more precise) in the single variable case, (Theorem (A), §8.3).

The reader will note that the problem treated in §11.5 is an example of multivariate analysis of variance, and is the multivariate analogue of the problem treated in §9.1.

To illustrate how the extension would be made in a randomized block layout with r' rows and s' columns, let us consider the case in which y has two components y_1 and y_2. Let y_{1ij} and y_{2ij} be the values of y_1 and y_2 corresponding to the i-th row and j-th column of our layout, $i=1,2,\ldots,r'$, $j=1,2,\ldots,s'$.

The distribution assumption for the y_{1ij} and y_{2ij} is that $(y_{1ij}-m_1-R_{1i}-C_{1j})$ and $(y_{2ij}-m_2-R_{2i}-C_{2j})$, $(\sum_{i=1}^{r'}R_{1i} = \sum_{j=1}^{s'}C_{1j} = \sum_{i=1}^{r'}R_{2i} = \sum_{j=1}^{s'}C_{2j} = 0)$ are jointly distributed according to a normal bivariate law with zero means and variance-covariance matrix

$$\left\| \begin{matrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{matrix} \right\| .$$

Now suppose we wish to test the hypothesis that the "column effects" are zero for both y_1 and y_2. This hypothesis may be specified as follows:

(a)
$$\Omega:\begin{cases} -\infty < m_1,\ m_2,\ R_{1i},\ R_{2i},\ C_{1j},\ C_{2j} < \infty, \\ (i=1,2,\ldots,r'),\quad (j=1,2,\ldots,s'), \quad \left\|\begin{matrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{matrix}\right\| \text{ positive definite.} \\ \sum_i R_{1i} = \sum_i R_{2i} = \sum_j C_{1j} = \sum_j C_{2j} = 0. \end{cases}$$
$$\omega:\begin{cases} \text{The subspace in } \Omega \text{ obtained by setting} \\ \text{each } C_{1j} \text{ and each } C_{2j} = 0. \end{cases}$$

This hypothesis, which may be denoted by $H[(C_{1j},C_{2j})=0]$, is clearly the 2-variable analogue of $H[(C_j)=0]$ in §9.2.

Let $\bar{y}_{11.}$, $\bar{y}_{1.j}$, $\bar{y}_1$, S_{11R}, S_{11C}, S_{11E} have meanings as functions of y_{111}, $y_{112},\ldots,y_{1r's'}$ similar to those of $\bar{y}_{i.}$, $\bar{y}_{.j}$, $\bar{y}$, S_R, S_C, S_E as functions of $y_{11},y_{12},\ldots,y_{rs}$. Let $\bar{y}_{21.}$, $\bar{y}_{2.j}$, $\bar{y}_2$, S_{22R}, S_{22C}, S_{22E} have similar meanings as functions of y_{211}, $y_{212},\ldots,y_{2r's'}$. Let

(b)
$$S_{12R} = \sum_{i,j}(\bar{y}_{11.}-\bar{y}_1)(\bar{y}_{21.}-\bar{y}_2),$$
$$S_{12C} = \sum_{i,j}(\bar{y}_{1.j}-\bar{y}_1)(\bar{y}_{2.j}-\bar{y}_2),$$
$$S_{12E} = \sum_{i,j}(y_{11j}-\bar{y}_{11.}-\bar{y}_{1.j}+\bar{y}_1)(y_{21j}-\bar{y}_{21.}-\bar{y}_{2.j}+\bar{y}_2).$$

It may be verified that the likelihood ratio λ for testing the hypothesis $H[(C_{1j},C_{2j})=0]$ is given by $U_o^{n/2}$, where

(c)
$$U_o = \frac{\left|\begin{matrix} S_{11E} & S_{12E} \\ S_{12E} & S_{22E} \end{matrix}\right|}{\left|\begin{matrix} S_{11E}+S_{11C} & S_{12E}+S_{12C} \\ S_{21E}+S_{21C} & S_{22E}+S_{22C} \end{matrix}\right|}.$$

It follows from Theorem (D), §11.7, that the h-th moment of U_o when $H[(C_{1j},C_{2j})=0)$ is true, the special case of (x), §11.7, obtained by setting $s=2$, $k=r'+s'-1$, $r = r'$, $n = r's'$, i. e.

(d)
$$E(U_o^h) = \frac{\Gamma(\frac{r'(s'-1)}{2})\Gamma(\frac{r'(s'-1)-1}{2})\Gamma(\frac{(r'-1)(s'-1)}{2}+h)\Gamma(\frac{(r'-1)(s'-1)-1}{2}+h)}{\Gamma(\frac{r'(s'-1)}{2}+h)\Gamma(\frac{r'(s'-1)-1}{2}+h)\Gamma(\frac{(r'-1)(s'-1)}{2})\Gamma(\frac{(r'-1)(s'-1)-1}{2})},$$

using formula (r) in §11.5, this reduces to

(e)
$$E(U_o^h) = \frac{\Gamma(r'(s'-1)-1)\Gamma((r'-1)(s'-1)-1+2h)}{\Gamma(r'(s'-1)-1+2h)\Gamma((r'-1)(s'-1)-1)},$$

from which we can easily obtain the distribution of U_0 by the method used in deriving the distribution of Z in (u), §11.5.

The extension of the hypothesis specified in (a) and the corresponding U_0 to the case in which y has several components, say $y_1, y_2, \ldots, y_s$ is immediate. Similar results hold for testing the hypothesis that "row effects" are zero.

We cannot go into further details. The illustration given above will perhaps indicate how Theorem (D) can be used as a basis of significance tests for multivariate analysis of variance problems arising in three-way layouts, Latin squares, Graeco-Latin squares, etc.

11.9 Principal Components of a Total Variance

Suppose $x_1, x_2, \ldots, x_k$ are distributed according to the normal multivariate* law (a) in §11.3. The probability density is constant on each member of the family or nest of k-dimensional ellipsoids

$$\text{(a)} \qquad \sum_{i,j=1}^{k} A_{ij}(x_i - a_i)(x_j - a_j) = C,$$

where $0 < C < \infty$. The ellipsoids in this family all have the same center $(a_1, a_2, \ldots, a_k)$ and are similarly situated with respect to their principal axes, that is, their longest axis lie on the same line, their second longest axis lie on the same line, etc., (assuming each has a longest, second longest,...,axis).

Our problem here is to determine the directions of the various principal axes, and the relative lengths of the principal axes for any given ellipsoid in the family (the ratios of lengths are, in fact, the same for each member of the family). We must first define analytically what is meant by principal axes. For convenience, we make the following translation of coordinates

$$\text{(b)} \qquad x_i - a_i = y_i \qquad i=1,2,\ldots,k.$$

The equation (a) now becomes

$$\text{(c)} \qquad \sum_{i,j=1}^{k} A_{ij} y_i y_j = C.$$

---

*The theory of principal axes and principal components as discussed in this section (including no sampling theory) can be carried through formally without assuming that the random variables $x_1, x_2, \ldots, x_k$ are distributed according to a normal multivariate law. However, this law is of sufficient interest to justify our use of it throughout the section. Some sampling theory of principal components under the assumption of normality will be presented in §11.11.

If $P:(y_1,y_2,\ldots,y_k)$ represents any point on this ellipsoid, then the squared distance D^2 between P and the center O is $\sum_1^k y_i^2$. Now if we allow P to move continuously over the ellipsoid, there will, in general, be 2k points at which the rate of change of D^2 with respect to the coordinates of P will be zero, i. e. there are 2k extrema for D^2 under these conditions. These points occur in pairs, the points in each pair being symmetrically located with respect to the center. The k line segments connecting the points in each pair are called <u>principal axes</u>. In the case of two variables, i. e. k = 2, our ellipsoids are simply ellipses, and the principal axes are the major and minor axes. We shall determine the points in the k-variate case and show that the principal axes are mutually perpendicular.

It follows from §4.7 that the problem of finding the extrema of D^2 for variations of P over (c) is equivalent to finding unrestricted extrema of the function

$$\phi = \sum_1^k y_i^2 + \lambda(C - \sum_{i,j=1}^k A_{ij}y_iy_j) \tag{d}$$

for variations of the y_i and λ. Following the Lagrange method, we must have

$$\frac{\partial\phi}{\partial y_i} = 0, \qquad (i=1,2,\ldots,k) \tag{e}$$

and also equation (c) satisfied. Performing the differentiations in (e), we obtain the following equations

$$y_i - \lambda \sum_{j=1}^k A_{ij}y_j = 0, \qquad (i=1,2,\ldots,k). \tag{f}$$

Suppose we multiply the i-th equation by A^{ih}, i=1,2,...,k, and sum with respect to i. We have

$$\sum_{i=1}^k A^{ih}y_i - \lambda \sum_{i,j=1}^k A_{ij}A^{ih}y_j = 0. \tag{g}$$

Since $\sum_1^k A_{ij}A^{ih} = 1$, if j = h, and 0, if $j \neq h$, (g) reduces so that it may be written as

$$-\lambda y_j + \sum_{i=1}^k A^{ij}y_i = 0. \tag{h}$$

Allowing j to take values 1,2,...,k, it is now clear that equations (h) are equivalent to (f) for finding the extrema. In order that (h) have solutions other than zero, it is necessary for

$$\begin{vmatrix} A^{11}-\lambda & A^{12} & \cdots & A^{1k} \\ A^{21} & A^{22}-\lambda & \cdots & A^{2k} \\ \vdots & & & \\ A^{k1} & & \cdots & A^{kk}-\lambda \end{vmatrix} = 0. \tag{i}$$

This equation is a polynimial of degree k, usually called the characteristic equation of the matrix $||A^{ij}||$. It can be shown that the roots of (i) are all real*. If the roots are all distinct, let them be $\lambda_1,\lambda_2,\ldots,\lambda_k$. The direction of the principal axis corresponding to λ_g is given by substituting λ_g in (h) and solving** for the y_j, $j=1,2,\ldots,k$. Let the values of the y_i $(i=1,2,\ldots,k)$ corresponding to λ_g be y_{gi} $(i=1,2,\ldots,k)$ and let the direction cosines of the g-th principal axis be c_{gi} (defined by $c_{gi} = y_{gi}/\sqrt{\sum_i y_{gi}^2}$). Hence, we have from (f)

$$(j) \qquad y_{gi} - \lambda_g \sum_{j=1}^{k} A_{ij} y_{gj} = 0, \qquad (i=1,2,\ldots,k).$$

It is now clear that if y_{gi} are solutions of (j), then $-y_{gi}$ are solutions also. Multiplying the i-th equation by y_{gi} and summing with respect to i, we find

$$(k) \qquad \sum_{i=1}^{k} y_{gi}^2 - \lambda_g \sum_{i,j=1}^{k} A_{ij} y_{gi} y_{gj} = 0,$$

or

$$(l) \qquad \sum_{i=1}^{k} y_{gi}^2 - \lambda_g C = 0.$$

Therefore, the squared length of half the g-th principal axis is $\lambda_g C$. If we consider the h-th principal axis, $(g \neq h)$, we have

$$(m) \qquad y_{hi} - \lambda_h \sum_{j=1}^{k} A_{ij} y_{hj} = 0, \qquad (i=1,2,\ldots,k).$$

If the i-th equation in (j) be multiplied by y_{hi}/λ_g and summed with respect to i, and if the i-th equation in (m) be multiplied by y_{gi}/λ_h and summed with respect to i, we obtain, upon combining the two resulting equations

$$(n) \qquad \sum_{i=1}^{k} y_{gi} y_{hi}/\lambda_g = \sum_{i=1}^{k} y_{gi} y_{hj}/\lambda_h.$$

Since $\lambda_g \neq \lambda_h$, this equation implies that $\sum_{i=1}^{k} y_{gi} y_{hi} = 0$, which means that the g-th and h-th $(g \neq h)$ principal axes are perpendicular, i. e. all principal axes are mutually perpendicular.

Suppose we change to a new set of variables defined as follows

---

*See M. Bôcher, Introduction to Higher Algebra, MacMillan Co., (1929), p. 170.

**For an iterative method of solving the equations, together with a more detailed treatment of principal components than we can consider here, see H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components", Jour. of Educ. Psych., Vol. 24, (1933), pp. 417-441, 498-520.

(o) $$\sum_{i=1}^{k} c_{gi} y_i = z_g, \qquad (g=1,2,\ldots,k).$$

Multiplying the g-th equation by c_{gj} and using the fact that $\sum_{g=1}^{k} c_{gi}c_{gj} = \delta_{ij}$ (for a set of mutually orthogonal vectors) and summing with respect to g, we find

(p) $$y_j = \sum_{g=1}^{k} c_{gj} z_g.$$

Substituting in the equation of the ellipsoid (c), we have

(q) $$\sum_{g,h=1}^{k} \sum_{i,j=1}^{k} A_{ij} c_{gi} c_{hj} z_g z_h = C.$$

Now it follows from the argument leading to (n) that $\sum_{i,j=1}^{k} A_{ij} c_{gi} c_{hj} = 0$, (if $g \neq h$), and $1/\lambda_g$, (if $g=h$). Hence the equation of the ellipsoid in the new coordinates is

(r) $$\sum_{g=1}^{k} \frac{z_g^2}{\lambda_g} = C.$$

The Jacobian of the transformation (p) is $|c_{gi}|$ which has the value 1, as one can see by squaring the determinant. Hence, if the $(x_i - a_i)$ are distributed according to the normal multivariate law (a), §11.3, and since (p) transforms the quadratic form (a) into (r), then the z_g are independently distributed with variance λ_g. But from (o) we also have, by taking variances of both sides,

$$\sum_{i,j=1}^{k} A^{ij} c_{gi} c_{gj} = \lambda_g.$$

Summing with respect to g, and using the fact that $\sum_{g=1}^{k} c_{gi}c_{gj} = \delta_{ij}$, we have

$$\sum_{i=1}^{k} A^{ii} = \sum_{g=1}^{k} \lambda_g.$$

In other words the sum of the variances of the y_i $(i=1,2,\ldots,k)$ is equal to the sum of the variances of the z_g $(g=1,2,\ldots,k)$. $\lambda_1, \lambda_2, \ldots, \lambda_k$ are called <u>principal components</u> of the total variance. It will be observed that z_g is constant on $(n-1)$-dimensional planes perpendicular to the α-th principal axis, $g=1,2,\ldots,k$.

We may summarize in the following

<u>Theorem (A)</u>: <u>Let $y_1, y_2, \ldots, y_k$ be random variables distributed according to the normal multivariate distribution</u>

(s) $$\frac{\sqrt{A}}{(2\pi)^{k/2}} e^{-\frac{1}{2}\sum_{i,j=1}^{k} A_{ij} y_i y_j} dy_1 \ldots dy_k.$$

Let the roots of the characteristic equation $|A^{ij}-\lambda\delta_{ij}| = 0$ be $\lambda_1,\lambda_2,\ldots,\lambda_k$. Let c_{gi} $(i=1,2,\ldots,k)$ be the direction cosines of the g-th principal axis of

$$\text{(t)} \qquad \sum_{i,j=1}^{k} A_{ij}y_iy_j = C,$$

and let

$$\text{(u)} \qquad \sum_{i=1}^{k} c_{gi}y_i = z_g, \qquad (g=1,2,\ldots,k).$$

Then

(1) The direction cosines are given by

$$c_{gi} = y_{gi}\Big/\left[\sum_{i=1}^{k}(y_{gi})^2\right]^{\frac{1}{2}}$$

where the y_{gi} satisfy the equations

$$-\lambda_g y_{gi} + \sum_{j=1}^{k} A^{ij}y_{gj} = 0, \qquad (i=1,2,\ldots,k).$$

(2) The length of half the g-th principal axis is $\sqrt{\lambda_g C}$.

(3) The principal axes are mutually perpendicular.

(4) The transformation (u) transforms the probability element (s) into

$$\text{(v)} \qquad \frac{1}{(2\pi)^{k/2}\sqrt{\lambda_1.\lambda_2.\ldots.\lambda_k}}\, e^{-\frac{1}{2}\sum_{g=1}^{k} z_g^2/\lambda_g}\, dz_1\ldots dz_k.$$

the z_g being independently distributed.

(5) $\sum_{i=1}^{k} A^{ii} = \sum_{g=1}^{k} \lambda_g$, i. e. the sum of the variances of the y_i is equal to the sum of the variances of the z_g.

If two of the roots of (1) are equal, we would have an indeterminant situation with reference to two of the principal axes. In this case, there will be a two-dimensional space, i. e. plane, perpendicular to each of the remaining principal axes such that the intersection of this plane with (c) is a circle. Similar remarks can be made about higher multiplicities of roots.

As a simple example in multiplicity of roots, the reader will find it instructive to consider the case in which the variance of y_i $(i=1,2,\ldots,k)$ is σ^2 and the covariance between y_i and y_j is $\sigma^2\rho$. Equation (1) becomes

$$[\sigma^2(1-\rho)-\lambda]^{k-1}[\sigma^2(1+(k-1)\rho)-\lambda] = 0.$$

There are roots of two magnitudes, one being $\sigma^2(1-\rho)$ with multiplicity k-1; the other being $\sigma^2(1+(k-1)\rho)$ with multiplicity 1. It is convenient in this case to think of one

long principal axis (if $\rho>0$) and k-1 short ones all equal (although indeterminate in direction). If $\rho>0$, then it is clear that the long axis increases as k increases, while the short axes remain the same. Thus the variance of the z (which is a linear function of the y_i by transformation (u)) corresponding to the longest axis increases with k. This property of increasing variance of the linear function of several positively intercorrelated variables associated with the longest axis, is fundamental in the scientific construction of examinations, certain kinds of indices, etc. By continuity considerations one can verify that the property holds, roughly speaking, even when the variances (as well as the covariances) of the variables depart slightly from each other.

11.10 Canonical Correlation Theory

Let $x_1,x_2,\ldots,x_k$ be random variables divided in two sets $S_1:(x_1,x_2,\ldots,x_{k_1})$ and $S_2:(x_{k_1+1},\ldots,x_{k_1+k_2})$ $(k_1+k_2=k)$. We shall assume that $k_1 \leq k_2$. Let L_1 and L_2 be arbitrary linear functions of the two groups of variates, respectively, i. e.

$$\text{(a)} \qquad \begin{aligned} L_1 &= \sum_{i=1}^{k_1} l_{1i}x_i, \\ L_2 &= \sum_{p=k_1+1}^{k} l_{2p}x_p. \end{aligned}$$

The correlation coefficient between L_1 and L_2 (see §2.75) is given by

$$R_{12} = \frac{\sum_{i,p} A^{ip} l_{1i} l_{2p}}{\sqrt{\left(\sum_{i,j} A^{ij} l_{1i} l_{1j}\right)\left(\sum_{p,q} A^{pq} l_{2p} l_{2q}\right)}},$$

where i and j in the summations range over the values $1,2,\ldots,k$, while p and q range over the values $k_1+1, k_1+2,\ldots,k_1+k_2$. $||A^{ip}||$ is the covariance matrix between the variables in G_1 and those in G_2; $||A^{ij}||$ is the variance-coveriance matrix for variables in G_1; a similar meaning holding for $||A^{pq}||$.

Now suppose we consider the problem* of varying the l_{1i} and l_{2p} so as to maximize the correlation coefficient R_{12}, (actually to find extrema of R_{12}, among which there will be a maximum). Corresponding to any given solution of this problem say l^*_{1i}, l^*_{2p}, $(i=1,2,\ldots,k_1;\ p=k_1+1,\ldots,k_1+k_2)$ there are infinitely many solutions of the form al^*_{1i}, bl^*_{2p}, where a and b are any two constants of the same sign. To overcome this difficulty, it is sufficient to seek a solution for fixed values of the variances of L_1 and L_2, which,

*This problem was first considered by H. Hotelling, "Relations Between Two Sets of Variates" Biometrika, Vol. 28 (1936), pp. 322-377.

for convenience we may take as 1. This is equivalent to the determination of the extrema of R_{12} for variations of the l_{1i} and l_{2p}, subject to the conditions

$$\text{(c)} \qquad \sum_{i,j} A^{ij} l_{1i} l_{1j} = 1, \qquad \sum_{p,q} A^{pq} l_{2p} l_{2q} = 1.$$

By Lagrange's method this amounts to finding the extrema of the function

$$\text{(d)} \qquad \phi = \sum_{i,p} A^{ip} l_{1i} l_{2p} + \tfrac{\lambda}{2}\Big(1 - \sum_{i,j} A^{ij} l_{1i} l_{1j}\Big) + \tfrac{\mu}{2}\Big(1 - \sum_{p,q} A^{pq} l_{2p} l_{2q}\Big),$$

where λ and μ are divided by 2 for convenience. The l_{1i} and l_{2p} must satisfy the equations

$$\text{(e)} \qquad \frac{\partial \phi}{\partial l_{1i}} = 0, \qquad i=1,2,\ldots,k_1$$

$$\text{(f)} \qquad \frac{\partial \phi}{\partial l_{2p}} = 0, \qquad p=k_1+1,\ldots,k,$$

which are

$$\text{(g)} \qquad \sum_{p} A^{ip} l_{2p} - \lambda \sum_{j} A^{ij} l_{1j} = 0, \qquad i=1,2,\ldots,k_1,$$

$$\text{(h)} \qquad \sum_{i} A^{ip} l_{1i} - \mu \sum_{q} A^{pq} l_{2q} = 0, \qquad p=k_1+1,\ldots,k.$$

Multiplying (g) by l_{1i} and summing with respect to (i), then multiplying (h) by l_{2p}, summing with respect to p, and using (c), we obtain

$$\text{(i)} \qquad \lambda = \mu = \sum_{i,p} A^{ip} l_{1i} l_{2p}.$$

Therefore putting $\mu = \lambda$ in (h), we obtain a system of k linear homogeneous equations in the l_{1i} and l_{2p}. In order to have a solution not identically zero, the k-th order determinant of the equations (g) and (h) must vanish. That is

$$\text{(j)} \qquad \left| \begin{array}{c|c} -\lambda A^{ij} & A^{iq} \\ \hline A^{pj} & -\lambda A^{pq} \end{array} \right| = 0.$$

If we factor $\sqrt{A^{ii}}$ out of the i-th row and j-th column (i=1,2,...,k) and $\sqrt{A^{pp}}$ out of the p-th row and p-th column ($p=k_1+1,\ldots,k_1+k_2$), we find that (j) is equivalent to

$$\text{(k)} \qquad \left| \begin{array}{c|c} -\lambda \rho_{ij} & \rho_{iq} \\ \hline \rho_{pj} & -\lambda \rho_{pq} \end{array} \right| = 0,$$

where the ρ's are correlation coefficients, and $\rho_{ii} = \rho_{pp} = 1$. It can be shown* that the roots of (k) are all real, since the determinant (k) is the discriminant of $Q_1 - \lambda Q_2$, where Q_2 is the sum of the two quadratic forms in (c), and hence is positive definite. If the determinant in (k) is expanded by Laplace's method by the first k_1 columns (or rows) it is clear from the resulting expansion that (k) is a polynomial of degree k_1+k_2 in which the lowest power of λ is k_2-k_1. Hence by factoring out $\lambda^{k_2-k_1}$ we are left with a polynomial $f(\lambda)$ in λ of degree $2k_1$. Now any term in the Laplace expansion of (k) (by the first k_1 columns) is the product of a determinant of order k_1 and one of order k_2. If the first determinant has r rows chosen from the upper left hand block of (k), then the second determinant will contain $k_2 - (k_1-r)$ rows from the lower left hand block of (k). The product of these two determinants will therefore have $\lambda^{k_2-k_1+2r}$ as a factor. Therefore, by factoring $\lambda^{k_2-k_1}$ from each term in the Laplace expansion of (k), it is clear that the resulting polynomial, that is $f(\lambda)$, will contain only even powers of λ. Therefore, the $2k_1$ roots of $f(\lambda) = 0$ are real and of the form $\pm\lambda_1, \pm\lambda_2, \ldots, \pm\lambda_{k_1}$, where each λ is ≥ 0. Let $\lambda_1 = \rho_{(21)}$ and $-\lambda_1 = \rho_{(21-1)}$. Let l_{u11}, l_{u2p}, be the solutions of the equations (g) and (h) corresponding to the root $\rho_{(u)}$, $(u=1,2,\ldots,2k_1)$ and let $L_{(u)1}$, $L_{(u)2}$, be the values of L_1 and L_2 in (a) corresponding to the solutions l_{u11}, l_{u2p}. Remembering that $\mu = \lambda$, and inserting the u-th root $\rho_{(u)}$ in (g) and (h), we must have

$$(l) \qquad \sum_p A^{ip} l_{u2p} - \rho_{(u)} \sum_j A^{ij} l_{u1j} = 0, \qquad i=1,2,\ldots,k_1,$$

$$(m) \qquad \sum_p A^{ip} l_{u1i} - \rho_{(u)} \sum_q A^{pq} l_{u2q} = 0, \qquad p=k_1+1,\ldots,k_1+k_2.$$

Multiplying (l) by l_{u1i} and summing with respect to i, and making use of the fact that $\sum_j A^{ij} l_{u1j} l_{u1i} = 1$, we find

$$(n) \qquad \sum_{i,p} A^{ip} l_{u2p} l_{u1i} - \rho_{(u)} = 0.$$

The first term in (n) is simply the correlation coefficient between $L_{(u)1}$ and $L_{(u)2}$, and its value is $\rho_{(u)}$. If u is even, then the correlation between $L_{(u)1}$ and $L_{(u)2}$ is equal to that between $L_{(u+1)1}$ and $-L_{(u+1)2}$ (or $-L_{(u+1)1}$ and $L_{(u+1)2}$). It can be easily verified that the correlation between $L_{(2i)1}$ and $L_{(2j)2}$ $(i \neq j)$ is zero. Hotelling has called $L_{(u)1}$ and $L_{(u)2}$ the u-th canonical variates, and $\rho_{(u)}$ the canonical correlation coefficient between the canonical variates $L_{(u)1}$ and $L_{(u)2}$. Hence, the canonical correlations and therefore the roots

*M. Bôcher, loc. cit., p. 170.

of the equation (k) lie on the interval (-1,+1). If there exists a single largest root, it is the one such that when it is substituted in (g) and (h) we obtain solutions (i. e. values of l_{11} and l_{2p}), which, used in (a), will give the linear functions having maximum correlation. For further details on canonical correlation theory, the reader is referred to Hotelling's paper.

We may summarize our results in the following

Theorem (A): Let $S_1:(x_1,x_2,\ldots,x_{k_1})$ and $S_2:(x_{k_1+1},\ldots,x_k)$ be two sets of random variables where $k = k_1+k_2$ $(k_1 \leqq k_2)$. Let L_1 and L_2, as defined in (a), be linear functions of the variables in S_1 and S_2, respectively, such that the variances of L_1 and L_2 are unity. Let R_{12} be the correlation coefficient between L_1 and L_2. Then

(1) There are at most $2k_1$ distinct extrema of R_{12} for variations of the l_{11} and l_{21} in L_1 and L_2;

(2) These extrema correspond to the $2k_1$ roots of equation (k), which lie on the interval (-1,+1) and are symmetrically spaced with respect to the origin.

(3) The value of R_{12} corresponding to the u-th root $\rho_{(u)}$ of (k) is equal in value to $\rho_{(u)}$ itself (the u-th canonical correlation coefficient).

(4) The canonical correlation coefficient between the two canonical variates corresponding to any two numerically different values of $\rho_{(u)}$ is zero.

The reader should note that no assumptions have been made about the distribution function of the two sets of random variables, S_1 and S_2. We are able to maintain this degree of generality as long as we are considering canonical correlation theory of populations. However, the statistical value of this theory may be questionable if the distributions of the x's in G_1 and G_2 departs radically from the normal multivariate law. Again in studying sampling theory of canonical correlations, progress has been made only for the case of sampling from normal multivariate populations. Some of the sampling results are given in §11.11.

11.11 The Sampling Theory of the Roots of Certain Determinantal Equations

In the treatment of the theory of principal components (§11.9) and of the canonical correlation theory (§11.10), it was found that the roots of certain determinantal equations in which the matrices are variance-covariance matrices, played fundamental roles. In testing hypotheses concerning principal components, canonical correlations and allied topics, we are interested in the roots of the analogous equations in which the matrices are sample variance-covariance matrices. In the following sections, we shall derive the

distributions of the roots of several sample determinantal equations* when the samples are drawn from certain special multivariate normal populations. The distribution theory of the roots for more general assumptions has not yet been developed.

11.111 Characteristic Roots of One Sample Variance-covariance Matrix.

Let us consider a sample $O_n:(x_{i\alpha};\ i=1,2,\ldots,k;\ \alpha=1,2,\ldots,n)$ from a normal multivariate population, whose variance-covariance matrix has one root λ of multiplicity k. The variance-covariance matrix of the population is then of the form

$$\left\| \begin{array}{cccc} \lambda & 0 & \ldots & 0 \\ 0 & \lambda & \ldots & 0 \\ \cdot & \cdot & \ldots & \cdot \\ 0 & 0 & \ldots & \lambda \end{array} \right\|$$

and its inverse is

$$\left\| \begin{array}{cccc} 1/\lambda & 0 & \ldots & 0 \\ 0 & 1/\lambda & \ldots & 0 \\ \cdot & \cdot & \ldots & \cdot \\ 0 & 0 & \ldots & 1/\lambda \end{array} \right\|$$

The p. d. f. of this population is

$$\text{(a)} \qquad \frac{1}{(2\pi\lambda)^{k/2}}\, e^{-\frac{1}{2\lambda}\sum_{i=1}^{k}(x_i-a_i)^2}.$$

Let $a_{ij}=\sum_{\alpha=1}^{n}(x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j)$, where $\bar{x}_i=\frac{1}{n}\sum_{\alpha=1}^{n}x_{i\alpha}$. The a_{ij} are distributed according to $W_{n-1,k}(a_{ij};\frac{1}{\lambda}\delta_{ij})$, where $\delta_{ij}=1$, if i=j, and = 0, if $i\neq j$. We are interested in finding the distribution of the roots of

$$\text{(b)} \qquad \left| \begin{array}{cccc} a_{11}-l & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22}-l & \ldots & a_{2k} \\ \cdot & \cdot & \ldots & \cdot \\ a_{k1} & a_{k2} & \ldots & a_{kk}-l \end{array} \right| = 0,$$

* These distributions and their derivations were first published in the papers by R. A. Fisher, "The Sampling Distribution of Some Statistics Obtained from Non-linear Equations", Annals of Eugenics, Vol. 9 (1939) pp. 238-249, and by P. L. Hsu, "On the Distribution of Roots of Certain Determinantal Equations", Annals of Eugenics, Vol. 9 (1939), pp. 250-258. The derivations used in this section were developed by A. M. Mood (unpublished).

which is analogous to (1) §11.9. For a geometrical interpretation of these roots, the reader is referred to §11.9.

In §11.9, it was shown that for a matrix $||A_{ij}||$ there is a set of numbers c_{gi} $(i,g=1,2,\ldots,k)$ (direction cosines of the principal axes of the family of ellipsoids $\sum_{i,j=1}^{k} A_{ij}y_iy_j = C$) such that the transformation

$$\sum_{i=1}^{k} c_{gi}y_i = z_g \qquad (g=1,2,\ldots,k)$$

will yield

(c) $$\sum_{i,j=1}^{k} A_{ij}y_iy_j = \sum_{g,h=1}^{k}\ \sum_{i,j=1}^{k} A_{ij}c_{gi}c_{hj}z_gz_h = \sum_{g=1}^{k}\frac{z_g^2}{\lambda_g},$$

where the λ_g are roots of $|A^{ij}-\lambda\delta_{ij}| = 0$. Expressing the z_g in terms of the y_i in the middle member of (c), we get

$$\sum_{i,j=1}^{k} A_{ij}y_iy_j = \sum_{i,j=1}^{k}\ \sum_{g=1}^{k}\frac{c_{gi}c_{gj}}{\lambda_g}y_iy_j\ .$$

Hence,

$$A_{ij} = \sum_{g=1}^{k}\frac{c_{gi}c_{gj}}{\lambda_g}, \qquad (i,j=1,2,\ldots,k).$$

In a similar manner we can find numbers γ_{ih} $(i,h=1,2,\ldots,k)$ to express a_{ij} as

(d) $$a_{ij} = \sum_{h=1}^{k}\gamma_{ih}\gamma_{jh}l_h, \qquad (i,j=1,2,\ldots,k)$$

where the l_h are the roots of (b) and the γ_{ih} are elements of an orthogonal matrix $||\gamma_{ih}||$; that is $\sum_{h=1}^{k}\gamma_{ih}\gamma_{jh} = \delta_{ij}$ and $\sum_{h=1}^{k}\gamma_{hi}\gamma_{hj} = \delta_{ij}$. The l_h and the γ_{ih} depend only on the a_{ij}. We can get the simultaneous distribution of l_h and γ_{ih} by substituting (d) in $w_{n-1,k}(a_{ij};\frac{1}{\lambda}\delta_{ij})$ and multiplying by the Jacobian of the transformation (d). Ordering the l_h so that $l_1 \geq l_2 \geq \ldots \geq l_k \geq 0$, the Jacobian is

(e) $$(l_1-l_2)(l_1-l_3)\ldots(l_1-l_k)(l_2-l_3)\ldots(l_{k-1}-l_k)\phi(\gamma_{ih}),$$

where $\phi(\gamma_{ih})$ is a function of the γ_{ih} only, not involving l_h. This can be verified in the following way. It is clear from (d) that the Jacobian will be a polynomial in the l_h; in fact it will be a polynomial of degree $\frac{k(k-1)}{2}$, for there are $\frac{k(k-1)}{2}$ independent elements in $||\gamma_{ij}||$. If $l_i = l_j$ $(i\neq j)$, the transformation (d) will not be uniquely determined, and hence the Jacobian will be zero since when a transformation is not (locally) unique, the Jacobian is zero. This fact implies that we can factor out terms (l_i-l_j) $(i\neq j)$. There are $\frac{k(k-1)}{2}$ such terms, and when they have been factored out, what

remains is independent of the l_h since the Jacobian is a polynomial of degree $\frac{k(k-1)}{2}$.

Noting that

$$|a_{ij}| = |\sum_{h=1}^{k}\gamma_{ih}l_h\gamma_{jh}| = |\sum_{g,h=1}^{k}\gamma_{ig}\delta_{gh}l_h\gamma_{jh}| = |\gamma_{ig}|\cdot|\delta_{gh}l_h|\cdot|\gamma_{jh}| = |\gamma_{ij}|^2\prod_{h=1}^{k}l_h$$

and that $\sum_{i=1}^{k}a_{ii} = \sum_{i=1}^{k}\sum_{h=1}^{k}\gamma_{ih}l_h\gamma_{ih} = \sum_{h=1}^{k}l_h$, we can write the distribution of the l_h and the γ_{ih} as

$$(f)\qquad \frac{(\prod_{i=1}^{k}l_i)^{\frac{n-k-2}{2}}\, e^{-\frac{1}{2\lambda}\sum_{i=1}^{k}l_i}\prod_{i<j}(l_i-l_j)}{(2\lambda)^{\frac{(n-1)k}{2}}\,\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}\Gamma(\frac{n-i}{2})}\,\phi(\gamma_{ij}).$$

To derive the distribution of the l_h alone, we integrate (f) with respect to the γ_{ij} over the space of the γ_{ij} for which $||\gamma_{ij}||$ is orthogonal, obtaining

$$(g)\qquad K(\prod_{i=1}^{k}l_i)^{\frac{n-k-2}{2}}\, e^{-\frac{1}{2\lambda}\sum_{i=1}^{k}l_i}\prod_{i<j}(l_i-l_j).$$

The constant K is determined by the condition that the integral of (g) over the space R of the l_h is unity. To find K, let us first define

$$(h)\qquad \phi(r) = \int_R(\prod_{i=1}^{k}l_i)^r\, e^{-\frac{1}{2\lambda}\sum_{i=1}^{k}l_i}\prod_{i<j}(l_i-l_j)\prod_{i=1}^{k}dl_i.$$

We note that

$$(i)\qquad K = \frac{1}{\phi(\frac{n-k-2}{2})}.$$

Since $\prod_{i=1}^{k}l_i = |a_{ij}|$, we have

$$(j)\qquad E(|a_{ij}|^s) = \frac{\phi(\frac{n-k-2}{2}+s)}{\phi(\frac{n-k-2}{2})},$$

but from the Wishart distribution (see (h) in §11.4), we find that

$$(k)\qquad E(|a_{ij}|^s) = (2\lambda)^{ks}\prod_{i=1}^{k}\frac{\Gamma(\frac{n-i}{2}+s)}{\Gamma(\frac{n-i}{2})}.$$

Equating (j) to (k) and setting n = k+2, we get

$$\frac{\phi(s)}{\phi(0)} = (2\lambda)^{ks}\prod_{i=1}^{k}\frac{\Gamma(\frac{k+2-i}{2}+s)}{\Gamma(\frac{k+2-i}{2})}. \tag{l}$$

It remains to evaluate

$$\phi(0) = \int_R e^{-\frac{1}{2\lambda}\sum_{i=1}^{k} l_h}\prod_{i<j}(l_i-l_j)\prod_{i=1}^{k} dl_i.$$

It can be verified that

$$\int_0^\infty\int_{x_{k-1}}^\infty\cdots\int_{x_2}^\infty e^{-\sum_{i=1}^{k} x_i}\prod_{i<j}(x_i-x_j)\prod_{i=1}^{k} dx_i = \frac{\prod_{i=1}^{k}\Gamma(k+1-i)}{2^{\frac{k(k-1)}{2}}}$$

Making use of (r) in §11.5, the right hand side may be written as

$$\frac{\prod_{i=1}^{k}\left|\Gamma(\frac{k+1-i}{2})\Gamma(\frac{k+2-i}{2})\right|}{\pi^{k/2}}.$$

Hence

$$\phi(0) = \frac{\prod_{i=1}^{k}\left|\Gamma(\frac{k+1-i}{2})\Gamma(\frac{k+2-i}{2})\right|}{\pi^{k/2}}(2\lambda)^{\frac{k(k+1)}{2}}.$$

Using this result and equations (i) and (l), we find that

$$K = \frac{\pi^{k/2}}{(2\lambda)^{\frac{k(n-1)}{2}}\prod_{i=1}^{k}\left|\Gamma(\frac{n-i}{2})\Gamma(\frac{k+1-i}{2})\right|}.$$

Substituting in (g), we finally obtain as the distribution element of the characteristic roots of (b)

$$\frac{\pi^{\frac{k}{2}}(\prod_{i=1}^{k} l_i)^{\frac{n-k-2}{2}} e^{-\frac{1}{2\lambda}\sum_{i=1}^{k} l_i}\prod_{i<j}(l_i-l_j)}{(2\lambda)^{\frac{k(n-1)}{2}}\prod_{i=1}^{k}\left|\Gamma(\frac{n-i}{2})\Gamma(\frac{k+1-i}{2})\right|}\prod_{i=1}^{k} dl_i. \tag{m}$$

It can be shown fairly readily by making appropriate orthogonal transformations, that if the sample $O_n:(x_{i\alpha};\ i=1,2,\ldots,k;\ \alpha=1,2,\ldots,n)$ is from the normal multivariate population (a) in §11.3 for the case in which the characteristic roots of the matrix

$|A^{ij}-\lambda\delta_{ij}| = 0$ are all equal to λ, say, then the characteristic roots of (b) are also distributed according to (m).

We may summarize in the following

Theorem (A): Let $O_n:(x_{i\alpha};\ i=1,2,\ldots,k;\ \alpha=1,2,\ldots,n)$ be a sample from a normal multivariate population for which the characteristic roots of the variance-covariance matrix are equal to λ. Let a_{ij} $(i,j=1,2,\ldots,k)$ be the second order sample product sums as defined below (a). Let $l_1,l_2,\ldots,l_k$ be the roots (in descending order of magnitude) of $|a_{ij}-l\delta_{ij}| = 0$. The joint probability element of the l_i $(i=1,2,\ldots,k)$ is given by (m).

11.112 Characteristic Roots of the Difference of Two Sample Variance-covariance Matrices.

Let us consider two samples $O_{n_1}:(x^1_{i\alpha},\ i=1,2,\ldots,k;\ \alpha=1,2,\ldots,n_1)$ and $O_{n_2}:(x^2_{i\alpha},\ i=1,2,,..,k;\ \alpha=1,2,\ldots,n_2)$ $(n_1>k,\ n_2>k)$ drawn from the same normal multivariate population

$$\text{(a)}\qquad \frac{\sqrt{A}}{(2\pi)^{k/2}}\,e^{-\frac{1}{2}\sum_{i,j=1}^{k}A_{ij}(x_i-a_i)(x_j-a_j)}.$$

Let $a^t_{ij} = \sum_{\alpha=1}^{n_t}(x^t_{i\alpha}-\bar{x}^t_i)(x^t_{j\alpha}-\bar{x}^t_j)$, $(t=1,2,)$. In this section, we shall derive the distribution of the roots of

$$\text{(b)}\qquad |\theta(a^1_{ij}+a^2_{ij}) - a^2_{ij}| = 0.$$

In §11.9, we have seen that there is a linear transformation

$$\text{(c)}\qquad x_i - a_i = \sum_{g=1}^{k} c_{gi}z_g \qquad (i=1,2,\ldots,k)$$

such that

$$\text{(d)}\qquad \sum_{i,j=1}^{k}A_{ij}(x_i-a_i)(x_j-a_j) = \sum_{g=1}^{k}\frac{z_g^2}{\lambda_g}.$$

Now let

$$\frac{z_g}{\sqrt{\lambda_g}} = w_g, \qquad (g=1,2,\ldots,k),$$

i.e.

$$\text{(e)}\qquad x_i - a_i = \sum_{g=1}^{k}\sqrt{\lambda_g}c_{gi}w_g, \qquad (i=1,2,\ldots,k).$$

Then

$$\sum_{i,j=1}^{k} A_{ij}(x_i-a_i)(x_j-a_j) = \sum_{g=1}^{k} w_g^2.$$

The transformation (e) when performed on the sample values gives us

$$a_{ij}^t = \sum_{c=1}^{n_t} \sum_{g,h=1}^{k} \sqrt{\lambda_g} c_{gi}(w_{i\alpha}^t - \bar{w}_i^t)\sqrt{\lambda_h} c_{hj}(w_{j\alpha}^t - \bar{w}_j^t) = \sum_{g,h=1}^{k} b_{ij}^t d_{gi} d_{hj} \qquad (t=1,2)$$

where

$$b_{ij}^t = \sum_{\alpha=1}^{n_t} (w_{i\alpha}^t - \bar{w}_i^t)(w_{j\alpha}^t - \bar{w}_j^t),$$

$$d_{gi} = \sqrt{\lambda_g} c_{gi}.$$

Now equation (b) becomes

$$(f) \qquad |\theta(a_{ij}^1 + a_{ij}^2) - a_{ij}^2| = \Big| \sum_{g,h=1}^{k} d_{gi}[\theta(b_{ij}^1 + b_{ij}^2) - b_{ij}^2] d_{hj} \Big|$$

$$= |d_{gi}| \cdot |\theta(b_{ij}^1 + b_{ij}^2) - b_{ij}^2| \cdot |d_{hj}| = 0.$$

Clearly the roots of

$$|\theta(b_{ij}^1 + b_{ij}^2) - b_{ij}^2| = 0$$

are the same as those of equation (b). Note that the b_{ij}^t are functions of the $w_{i\alpha}^t$, such that each value of i, t, and α, $w_{i\alpha}^t$ is distributed according to a normal law with zero mean and unit variance, the $w_{i\alpha}^t$ being independently distributed. This shows that we lose no generality by assuming that $A_{ij} = 1$, if $i=j$, and $= 0$ if $i \neq j$.

Under this assumption, the a_{ij}^t have the distribution

$$(g) \qquad w_{n_t-1,k}(a_{ij}^t; \delta_{ij}) = \frac{|a_{ij}^t|^{\frac{n_t-k-2}{2}} e^{-\frac{1}{2}\sum_{i=1}^{k} a_{ii}^t}}{2^{\frac{k(n_t-1)}{2}} \pi^{\frac{k(k-1)}{4}} \prod_{i=1}^{k} \Gamma(\frac{n_t-i}{2})}, \qquad (t=1,2).$$

From a theorem in algebra* we know that there is a transformation of the a_{ij}^t such that $(i,j=1,2,\ldots,k)$

$$(h) \qquad a_{ij}^1 + a_{ij}^2 = \sum_{h=1}^{k} u_{ih} u_{jh}$$

$$a_{ij}^2 = \sum_{h=1}^{k} u_{ih} \theta_h u_{jh},$$

where θ_h are the roots of (b) (arranged, say, in descending order of magnitude). The u_{ih}

*See M. Bôcher, loc. cit. p. 171.

and the θ_h are functions of a_{ij}^1 and a_{ij}^2; hence, their distribution may be found by substituting in

$$(i) \qquad w_{n_1-1,k}(a_{ij}^1;\delta_{ij}) \cdot w_{n_2-1,k}(a_{ij}^2;\delta_{ij})$$

and multiplying by the Jacobian of the transformation (h). By following a procedure similar to that of §11.111, we can show that the Jacobian of (h) is

$$\frac{\partial(a_{ij}^1,\partial a_{ij}^2)}{\partial(\theta_h,\partial u_{ij})} = \prod_{i<j}(\theta_i-\theta_j)\cdot\ \psi(u_{ij}),$$

where $\psi(u_{ij})$ is a function of the u_{ij} independent of θ_h. Hence, the simultaneous distribution of the θ_h and u_{ij} is

$$(j) \qquad \frac{\left|\sum_{h=1}^{k} u_{ih}u_{jh} - \sum_{h=1}^{k} u_{ih}\theta_h u_{jh}\right|^{\frac{n_1-k-2}{2}} e^{-\frac{1}{2}\sum_{i=1}^{k}\sum_{h=1}^{k}(u_{ih}^2-u_{ih}^2\theta_h)}}{2^{\frac{k(n_1-1)}{2}}\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}\Gamma(\frac{n_1-1}{2})} \cdot \frac{\left|\sum_{h=1}^{k} u_{ih}\theta_h u_{jh}\right|^{\frac{n_2-k-2}{2}} e^{-\frac{1}{2}\sum_{i=1}^{k}\sum_{h=1}^{k}u_{ih}^2\theta_h}}{2^{\frac{k(n_2-1)}{2}}\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}(\frac{n_2-1}{2})} \prod_{i<j=1}^{k}(\theta_i-\theta_j)\psi(u_{ij}).$$

Noting that $\left|\sum_{h=1}^{k} u_{ih}(1-\theta_h)u_{jh}\right| = |u_{ih}| \cdot |(1-\theta_h)\delta_{gh}| \cdot |u_{jh}|$ and $\left|\sum_{h=1}^{k} u_{ih}\theta_h u_{jh}\right| = |u_{ih}| \cdot |\theta_h\delta_{gh}| \cdot |u_{jh}|$ we see that (j) factors into a function of the θ_i and a function of the u_{ij}

$$(k) \quad C\left[\prod_{i=1}^{k}(1-\theta_i)\right]^{\frac{n_1-k-2}{2}}\left[\prod_{h=1}^{k}\theta_i\right]^{\frac{n_2-k-2}{2}}\prod_{i<j}(\theta_i-\theta_j) \cdot |u_{ij}|^{n_1+n_2-2k-4} e^{-\frac{1}{2}\sum_{i,j=1}^{k}u_{ij}^2}\psi(u_{ij}),$$

where C is a constant. On integrating with respect to the u_{ij} we get the marginal distribution of the θ_i

$$(l) \qquad K\prod_{i=1}^{k}(1-\theta_i)^{\frac{n_1-k-2}{2}}\prod_{i=1}^{k}\theta_i^{\frac{n_2-k-2}{2}}\prod_{i<j}(\theta_i-\theta_j).$$

K is a constant to be determined so the integral of (l) over the range of the θ_i is unity. Following a procedure similar to that used in determining K in §11.111, we evaluate K in (l) and obtain as the distribution element of the θ_i

$$(m)\quad \pi^{k/2}\prod_{i=1}^{k}\frac{\Gamma(\frac{n_1+n_2-i-1}{2})}{\Gamma(\frac{n_1-i}{2})\Gamma(\frac{n_2-i}{2})\Gamma(\frac{k-i+1}{2})}\prod_{i=1}^{k}(1-\theta_i)^{\frac{n_1-k-2}{2}}\prod_{i=1}^{k}\theta_i^{\frac{n_2-k-2}{2}}\prod_{i<j=1}^{k}(\theta_i-\theta_j)\prod_{i=1}^{k}d\theta_i.$$

It should be emphasized that distribution (m) holds for the roots of (b) where the a_{ij}^1 and the a_{ij}^2 are any two sets of random variables distributed independently according to the Wishart distributions

$$(n)\qquad w_{n_1-1,k}(a_{ij}^1;A_{ij})\ ,\ w_{n_2-1,k}(a_{ij}^2;A_{ij}),$$

where n_1 and n_2 are both $> k$. In fact, we may summarize our results in

Theorem (A): Let a_{ij}^1 and a_{ij}^2 be independently distributed according to the Wishart distributions (n). Let $\theta_1,\theta_2,\ldots,\theta_k$ (in descending order) be the roots of the equation $|\theta(a_{ij}^1+a_{ij}^2)-a_{ij}^2| = 0$. Then the joint probability element of the θ_i $(i=1,2,\ldots,k)$ is given by (m), where the range of the θ's is $1\geq\theta_1\geq\theta_2\geq\ldots\geq\theta_k>0$.

11.113 Distribution of the Sample Canonical Correlations.

Corresponding to the population canonical correlations discussed in §11.10, there are canonical correlations of a sample. In this section, we shall determine the distribution of the sample canonical correlations when the smaller set of variates has a normal multivariate distribution independent of the other set.

Consider a sample $O_n:(x_{u\alpha};\ u=1,2,\ldots,k_1+k_2;\ \alpha=1,2,\ldots,n)$ from a population where the first k_1 variates have a normal distribution and the remaining k_2 variates are distributed independently of the first k_1 $(k_1\leq k_2)$. Let

$$a_{uv}=\sum_{\alpha=1}^{n}(x_{u\alpha}-\bar{x}_u)(x_{v\alpha}-\bar{x}_v),\qquad (u,v=1,2,\ldots,k_1+k_2).$$

The canonical correlations of the sample are defined as the roots of

$$(a)\qquad \left|\begin{array}{c|c} -la_{ij} & a_{iq} \\ \hline a_{pj} & -la_{pq} \end{array}\right| = 0,\qquad \begin{array}{l}(i,j=1,2,\ldots,k_1,\\ p,q=k_1+1,\ldots,k_1+k_2).\end{array}$$

Multiplying each of the first k_1 columns by 1 and then factoring 1 out of each of the last k_2 rows, we see that (a) is equivalent to

$$(b)\qquad \left|\begin{array}{c|c} -l^2a_{ij} & a_{iq} \\ \hline a_{pj} & -a_{pq} \end{array}\right| = 0,$$

except for a factor of $l^{k_2-k_1}$. Since we are not interested in the roots which are identically zero, we shall confine our attention to the roots of (b).

Let a^{pq} be the element corresponding to a_{pq} in the inverse of $||a_{pq}||$ $(p,q=k_1+1,\ldots,k_1+k_2)$. After multiplication on the left by

$$\text{(c)} \qquad \left|\begin{array}{c|c} \delta_{hi} & \sum_{r=k_1+1}^{k_1+k_2} a_{hr}a^{rp} \\ \hline 0 & a^{sp} \end{array}\right|$$

equation (b) becomes

$$\text{(d)} \qquad \left|\begin{array}{c|c} -l^2 a_{hj}+\sum_{p,r=k_1+1}^{k_1+k_2} a_{hr}a^{rp}a_{pj} & \sum_{i=1}^{k_1}\delta_{hi}a_{iq}-\sum_{p,r=k_1+1}^{k_1+k_2} a_{hr}a^{rp}a_{pq} \\ \hline \sum_{p=k_1+1}^{k_1+k_2} a^{sp}a_{pj} & -\sum_{p=k_1+1}^{k_1+k_2} a^{sp}a_{pq} \end{array}\right| = 0,$$

$$(h=1,2,\ldots,k_1;\ r,s=k_1+1,\ldots,k_1+k_2).$$

Since each member in the upper right hand block is 0 and since each element in the lower right hand block is δ_{sq}, (d) can be reduced to

$$\text{(e)} \qquad \left|-l^2 a_{hj}+\sum_{p,r=k_1+1}^{k_1+k_2} a_{hr}a^{rp}a_{pj}\right| = 0.$$

The roots of (e) (which are also roots of (b)) are the sample canonical correlation coefficients. Let the squared roots (in descending order) be $l_1^2, l_2^2, \ldots, l_{k_1}^2$. We observe that

$$\text{(f)} \qquad \sum_{p,r=k_1+1}^{k_1+k_2} a_{hr}a^{rp}a_{pj} = -\frac{\left|\begin{array}{c|c} 0 & a_{hr} \\ \hline a_{pj} & a_{pr} \end{array}\right|}{|\,a_{pr}\,|},$$

where, in the determinants on the right, h and j are fixed but $p,r = k_1+1,\ldots,k_1+k_2$. Let this value be b_{hj}. If we consider the $x_{p\alpha}$ $(\alpha = 1,2,\ldots,n;\ p = k_1+1,\ldots,k_1+k_2)$ fixed with $||a_{pq}||$ positive definite, then a_{ij} and b_{ij} are bilinear forms in $x_{i\alpha}$ and $x_{j\beta}$ $(\alpha,\beta = 1,2,\ldots,n;\ i,j = 1,2,\ldots,k_1)$; i. e., a_{ij} may be written as $\sum_{\alpha,\beta=1}^{n} G_{\alpha\beta}x_{i\alpha}x_{j\beta}$, where $||G_{\alpha\beta}||$ is of rank n-1, and b_{ij} may be written as $\sum_{\alpha,\beta=1}^{n} H_{\alpha\beta}x_{i\alpha}x_{j\beta}$, where $||H_{\alpha\beta}||$ is of rank k_2, $H_{\alpha\beta}$ being a function of the fixed $x_{p\alpha}$. a_{ij} and b_{ij} are, therefore, bilinear forms in the $x_{i\alpha}$ and $x_{j\beta}$ having matrices which do not depend on i and j. By Cochran's

Theorem, we know that there is a transformation which applied to the $x_{1\alpha}$ would make

$$a_{11} = \sum_{\alpha=1}^{n-1} y_{1\alpha}^2, \qquad b_{11} = \sum_{\alpha=1}^{k_2} y_{1\alpha}^2.$$

Applying this same transformation to each set we get

$$a_{ij} = \sum_{\alpha=1}^{n-1} y_{i\alpha} y_{j\alpha}, \qquad b_{ij} = \sum_{\alpha=1}^{k_2} y_{i\alpha} y_{j\alpha}.$$

The y's are normally and independently distributed with zero means and equal variances.

Thus (e) may be written in the form

$$|l(c_{ij}+b_{ij}) - b_{ij}| = 0,$$

where $c_{ij} = a_{ij}-b_{ij}$; the c_{ij} and b_{ij} being independently distributed according to Wishart distributions $w_{n-k_2-1,k_1}(c_{ij};B_{ij})$ and $w_{k_2,k_1}(b_{ij};B_{ij})$, where $||B_{ij}||$ is some positive definite matrix.

Therefore, it follows from the results of §11.113 that the square of the roots of (e) (i. e. the square of the canonical correlation coefficients) have the distribution (m) where $n_1 = n-k_2$, $k=k_1$ and $n_2 = k_2+1$. That is, the distribution is

$$(g) \qquad \pi^{\frac{k_1}{2}} \prod_{i=1}^{k_1} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-k_2-i}{2})\Gamma(\frac{k_2+1-i}{2})\Gamma(\frac{k_1+1-i}{2})} \prod_{i=1}^{k_1}(1-\theta_i)^{\frac{n-k_1-k_2-2}{2}} \prod_{i=1}^{k_1}\theta_i^{\frac{k_2-k_1-1}{2}} \prod_{i<j}(\theta_i-\theta_j) \prod_{i=1}^{k_1} d\theta_i,$$

where $\theta_i = l_i^2$, $i=1,2,\ldots,k_1$.

We may summarize our results in

Theorem (A): Let 0_n: $(x_{i\alpha}; i=1,2,\ldots,k_1+k_2; k_1 \leq k_2; \alpha=1,2,\ldots,n)$ be a sample from a population in which the first k_1 variates are distributed according to a normal multivariate distribution, but independently of the remaining k_2 variates (which may have any arbitrary distribution or may be "fixed" variates). Let a_{uv} $(u,v=1,2,\ldots,k_1+k_2)$ be the second order sample product sums as defined above (a), and let $l_1^2, l_2^2, \ldots, l_{k_1}^2$ be the squared roots (squared canonical correlation coefficients) of equation (b). The joint distribution element of the l_i^2, $(i=1,2,\ldots,k_1)$ is given by (g) where $\theta_i = l_i^2$, and where the range of the l^2's is such that $1 \geq l_1^2 \geq l_2^2 \geq \ldots \geq l_{k_1}^2 \geq 0$.

LITERATURE FOR SUPPLEMENTARY READING

1. American Standards Association: "Guide for Quality Control and Control Chart Method of Analyzing Data" (1941) and "Control Chart Method of Controlling Quality During Production" (1942), American Standards Association, New York.

2. Anderson, R. L.: "Distribution of the Serial Correlation Coefficient", Annals of Math. Stat., Vol. 13, (1942) pp. 1 - 13.

3. Bartlett, M. S.: "On the Theory of Statistical Regression", Proc. Royal Soc. of Edinburgh, Vol. 53 (1933), pp. 260 - 283.

4. Bartlett, M. S.: "The Effect of Non-Normality on the t Distribution, Proc. Camb. Phil. Soc., Vol. 31 (1935), pp. 223 - 231.

5. Battin, I. L.: "On the Problem of Multiple Matching", Annals of Math. Stat., Vol. 13 (1942), pp. 294 - 305.

6. Bôcher, M.: Introduction to Higher Algebra. MacMillan, New York (1907).

7. Bortkiewicz, L. von: Die Iterationen. Berlin, Springer, (1917).

8. Brown, George W.: "Reduction of a Certain Class of Statistical Hypotheses", Annals of Math. Stat., Vol. 11 (1940), pp. 254 - 270.

9. Camp, B. H.: "A New Generalization of Tchebycheff's Inequality", Bull. Amer. Math. Soc., Vol. 28 (1922), pp. 427 - 432.

10. Cochran, G. C.: "The Distribution of Quadratic Forms in a Normal System, with Applications to the Analysis of Covariance", Proc. Camb. Phil. Soc., Vol. 30 (1934), pp. 178 - 191.

11. Copeland, A. H.: "Point Set Theory Alplied to the Random Selection of the Digits of an Admissible Number", Amer. Jour. Math., Vol. 58 (1936), pp. 181 - 192.

12. Craig, C. C.: "On the Composition of Dependent Elementary Errors", Annals of Math., Vol. 33 (1932), pp. 184 - 206.

13. Craig, A. T.: "On the Distribution of Certain Statistics", Amer. Jour. Math., Vol. 54 (1932), pp. 353 - 366.

14. Cramer, H. and Wold, H.: "Some Theorems on Distribution Functions", Jour. London Math. Soc., Vol. 11 (1936), pp. 290 - 294.

15. Curtiss, J. H.: "On the Theory of Moment Generating Functions", Annals of Math. Stat., Vol. 13 (1942), pp. 430 - 433.

16. Daly, J. F.: "On the Unbiased Character of Likelihood Ratio Tests for Independence in Normal Systems", Annals of Math. Stat., Vol. 11 (1940), pp. 1 - 32.

17. Darmois, G.: Statistique Mathématique. Paris, Doin, 1928.

18. Deming, W. E., and Birge, R. T.: "On the Statistical Theory of Errors", Rev. Modern Phys., Vol. 6 (1934), pp. 122 - 161.

19. Dodd, E. L.: "Probability as Expressed by Asymptotic Limits of Pencils of Sequences" Bull. Amer. Math. Soc., Vol. 36, (1930), pp. 299 - 305.

20. Dodd, E. L.: "The Length of Cycles Which Result from the Graduation of Chance Elements", Annals of Math. Stat., Vol. 10 (1939), pp. 254 - 264.

21. Dodge, H. F., and Romig, H. G.: "A Method of Sampling Inspection", Bell System Tech. Jour., Vol. VIII, (1929).

22. Dodge, H. F., and Romig, H. G.: "Single Sampling and Double Sampling Inspection Tables", Bell System Tech. Jour., Vol. XX (1941).

23. Doob, J. L.: "Probability and Statistics", Trans. Amer. Math. Soc., Vol. 36 (1934), pp. 759 - 775.

24. Feller, Willy,: "On the Integral Equation of Renewal Theory", Annals of Math. Stat., Vol. 11 (1941), pp. 243 - 267.

25. Fertig, J. W.: "On a Method of Testing the Hypothesis that an Observed Sample of n Variables and of Size N has been drawn from a Specified Population of the Same Number of Variables", Annals of Math. Stat., Vol. 7 (1936), pp. 113 - 163.

26. Fertig, J. W.: "The Testing of Certain Hypotheses by means of Lambda Criteria with Particular Reference to Physiological Research", Biometric Bulletin, Vol. 1 (1936), pp. 45 - 82.

27. Fisher, R. A. and Yates, F.: Statistical Tables for Biological, Agricultural and Medical Research, London, Oliver and Boyd, 1938.

28. Fisher, R. A.: "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P., Jour. Roy. Stat. Soc., Vol. 85 (1922), pp. 87 - 94.

29. Fisher, R. A.: "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population", Biometrika, Vol. 10 (1915), pp. 507 - 521.

30. Fisher, R. A.: "On the Mathematical Foundations of Theoretical Statistics", Phil. Trans. Roy. Soc. London, Series A, Vol. 222 (1921), pp. 309 - 368.

31. Fisher, R. A.: "On a Distribution Yielding Error Functions of Several Well Known Statistics", Proc. Internat. Cong. of Math., Toronto (1924), pp. 805 - 813.

32. Fisher, R. A.: "The Theory of Statistical Estimation", Proc. Camb. Phil. Soc., Vol. 22 (1926), pp. 700 - 725.

33. Fisher, R. A.: "Applications of 'Student's' Distribution", Metron, Vol. 5 (1926), pp. 90 - 104.

34. Fisher, R. A.: "The General Sampling Distribution of the Multiple Correlation Coefficient", Proc. Roy. Soc. London, Series A, Vol. 121 (1928) pp. 654 - 673.

35. Fisher, R. A.: "Inverse Probability", Proc. Camb. Phil. Soc., Vol. 26 (1930), pp. 528 - 535.

36. Fisher, R. A.: "The Concepts of Inverse Probability and Fiducial Probability Referring to Unknown Parameters", Proc. Roy. Soc. London, Series A, Vol. 139 (1933), pp. 343 - 348.

37. Fisher, R. A.: "The Fiducial Argument in Statistical Inference", Annals of Eugenics, Vol. 6 (1935), pp. 391 - 398.

38. Fisher, R. A.: The Design of Experiments. London, Oliver and Boyd, 1935.

39. Fisher, R. A.: "The Sampling Distribution of Some Statistics Obtained from Non-Linear Equations", Annals of Eugenics, Vol. 9 (1939), pp. 238 - 249.

40. Fisher, R. A.: Statistical Methods for Research Workers. 8th Ed., London, Oliver and Boyd, 1941.

41. Fry, T. C.: Probability and its Engineering Uses. Van Nostrand Co., 1928.

42. Girshick, M. A.: "On the Sampling Theory of the Roots of Determinantal Equations", Annals of Math. Stat., Vol. 10 (1939), pp. 203 - 224.

43. Greville, T. N. E.: "The Frequency Distribution of a General Matching Problem", Annals of Math. Stat., Vol. 12 (1941), pp. 350 - 354.

44. Gumbel, E. J.: "Les Valewis Extrêmes des Distributions Statistiques", Annales de l'Institut H. Poincaré (1935).

45. Hamburger, H.: "Über eine Erweiterung des Stieltzesschen Momentenproblems", Math. Annalen, Vol. 81 (1920) pp. 235 - 319, and Vol. 82 (1921), pp. 120 - 165, 168 - 187.

46. Hotelling, H.: "The Generalization of Student's Ratio", Annals of Math. Stat., Vol. 2, (1931), pp. 359 - 378.

47. Hotelling, H.: "Analysis of a Complex of Statistical Variables into Principal Components", Jour. Ed. Psych., Vol. 24 (1933), pp. 417 - 441, pp. 498 - 520.

48. Hotelling, H.: "Relations between Two Sets of Variates", Biometrika, Vol. 28 (1936), pp. 321 - 377.

49. Hotelling, H.: "Experimental Determination of the Maximum of a Function", Annals of Math. Stat., Vol. 12 (1941), pp. 20 - 45.

50. Hsu, P. L.: "On the Distribution of Roots of Certain Determinantal Equations", Annals of Eugenics, Vol. 9 (1939) pp. 250 - 258.

51. Hsu, P. L.; "On Generalized Analysis of Variance", Biometrika, Vol. 31 (1940) pp. 221 - 237.

52. Ingham, A. E.: "An Integral which Occurs in Statistics", Proc. Camb. Phil. Soc., Vol. 29 (1933), pp. 270 - 276.

53. Irwin, J. O.: "Mathematical Theorems Involved in the Analysis of Variance", Jour. Roy. Stat. Soc., Vol. 94 (1931), pp. 284 - 300.

54. Irwin, J. O. and others: "Recent Advances in Mathematical Statistics", Jour. Roy. Stat. Soc., Vol. 95 (1932), Vol. 97 (1934), Vol. 99 (1936).

55. Kamke, E.: Einführung in die Wahrscheinlichkeitstheorie. Leipzig, Hirzel, 1932.

56. Kendall, M. G. and Smith, B. B.: "The Problem of m Rankings", Annals of Math. Stat Vol. 10 (1939), pp. 275 - 287.

57. Keynes, J. M.: A Treatise on Probability. MacMillan, London, 1921.

58. Kolodziejczyk, S.: "On an Important Class of Statistical Hypotheses", Biometrika, Vol. 27 (1935), pp. 161 - 190.

59. Koopman, B. O.: "On Distributions Admitting a Sufficient Statistic", Trans. Amer. Math. Soc., Vol. 39 (1936), pp. 399 - 409.

60. Koopmans, T.: On Modern Sampling Theory. Lectures delivered at Oslo, 1935, (unpublished).

61. Koopmans, T.: "Serial Correlation and Quadratic Forms in Normal Variables", Annals of Math. Stat., Vol. 13, (1942), pp. 14 - 33.

62. Kullback, S.: "An Application of Characteristic Functions to the Distribution Problem in Statistics", Annals of Math. Stat., Vol. 5 (1934), pp. 264 - 307.

63. Lawley, D. N.: "A Generalization of Fisher's z", Biometrika, Vol. 30 (1938), pp. 180 - 187.

64. Lévy, D.: Theorie de L'addition des Variables Aleatoires. (Monographies des probabilities, fasc. 1) Gauthier, 1937.

65. Lotka, Alfred J.: "A Contribution to the Theory of Self-Renewing Aggregates, with Special Reference to Industrial Replacement", Annals of Math. Stat., Vol. 10 (1939), pp. 1 - 25.

66. Madow, W. G.: "Contributions to the Theory of Multivariate Statistical Analysis", Trans. Amer. Math. Soc., Vol. 44 (1938), pp. 454 - 495.

67. Mises, R. von: Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und Theoretischen Physik., Leipzig, Deuticke, 1931.

68. Mood, A. M.: "The Distribution Theory of Runs", Annals of Math. Stat., Vol. 11 (1940), pp. 367 - 392.

69. Mosteller, Frederick, "Note on an Application of Runs to Quality Control Charts", Annals of Math. Stat., Vol. 12, (1941) pp. 228 - 232.

70. Neumann, J. von: "Distribution of the Ratio of the Mean Square Successive Difference to the Variance", Annals of Math. Stat., Vol. 12 (1941), pp. 367 - 395.

71. Neyman, J. and Pearson, E. S.: "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference", Biometrika, Vol. 20 A (1928) pp. 175 - 240, pp. 263 - 294.

72. Neyman, J. and Pearson, E. S.: "On the Problem of the Most Efficient Tests of Statistical Hypotheses", Phil. Trans. Roy. Soc., London, Ser. A, Vol. 231 (1933) p. 289.

73. Neyman, J. and Pearson, E. S.: "The Testing of Statistical Hypotheses in Relation to Probabilities a priori", Proc. Camb. Phil. Soc., Vol. 29 (1933), pp. 492 - 510.

74. Neyman, J. and Pearson, E. S.: Statistical Research Memoirs. University College, London, Vol. 1 (1936), Vol. 2 (1937).

75. Neyman, J.: "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection", Jour. Roy. Stat. Soc., Vol. 97 (1934), pp. 558 - 625.

76. Neyman, J.: "Su Un Teorema Concernete le Cosiddette Statistiche Sufficienti", Giornale dell' Instituto Italiano degli Attuari, Vol. 6 (1934), pp. 320 - 334.

77. Neyman, J.: "Outline of a Theory of Statistical Estimation based on the Classical Theory of Probability", Phil. Trans. Roy. Soc., London, Ser. A, Vol. 236 (1937) pp. 333 - 380.

78. Perron, O.: Die Lehre von der Kettenbrüchen. Leipzig, Teubner, 1929.

79. Pearson K.: "On the Criterion that a Given Set of Deviations from the Probable in the Case of Correlated Variables is Such that It Can Reasonably be Supposed to have Arisen from Random Sampling", Phil. Mag. 5th Ser., Vol. 50 (1900), pp. 157 - 175.

80. Pearson, K.: Tables for Statisticians and Biometricians. Cambridge University Press, 1914.

81. Pearson, K.: Tables of the Incomplete Gamma Function. Cambridge University Press, 1922.

82. Pearson, K.: Tables of the Incomplete Beta Function. Cambridge University Press, 1932.

83. Pomey, J. B.: Calcul des Probabilities. Paris, Gauthier-Villars, 1936.

84. Reichenbach, H.: Wahrscheinlichkeitslehre. Leiden, Sijthoff, 1935.

85. Rider, P. R.: "A Survey of the Theory of Small Samples", Annals of Math., Vol. 31, (1930), pp. 577 - 628.

86. Rietz, H. L.: Mathematical Statistics. Open Court Publishing Co., Chicago, 1927.

87. Sasuly, M.: Trend Analysis in Statistics. The Brookings Institution, Washington, 1934.

88. Scheffé, H.: "On the ratio of the variances of two normal populations", Annals of Math. Stat., Vol. 13 (1942), pp. 371 - 388.

89. Shewhart, W. A.: Economic Control of Quality of Manufactured Product. Van Nostrand, 1931.

90. Shewhart, W. A.: Statistical Method from the Viewpoint of Quality Control. U. S. Department of Agriculture, Washington, 1939.

91. Smirnoff, V. I.: "Sur les Écarts de la Courbe de Distribution Empirique", Recueil Mathématique, Moscow, Vol. 6 (1939) pp. 25 - 26.

92. Snedecor, G. W.: Calculation and Interpretation of Analysis of Variance and Covariance. Collegiate Press, Ames, Iowa, 1934.

93. Sommerville, D. M. Y.: An Introduction to the Geometry of N Dimensions. London, Methuen, (1929).

94. Stevens, W. L.: "Distribution of Groups in a Sequence of Alternatives", Annals of Eugenics, Vol. IX (1939).

95. "Student": "The Probable Error of a Mean". Biometrika, Vol. 6 (1908), pp. 1 - 25.

96. Swed, Frieda S., and Eisenhart, C.: "Tables for Testing Randomness of Grouping in a Sequence of Alternatives", Annals of Math. Stat., Vol. 14 (1943).

97. Wald A. and Wolfowitz, J.: "On a Test of Whether Two Samples are from the Same Population", Annals of Math., Stat., Vol. 11, (1940), pp. 147 - 162.

98. Wald, A.: "Contributions to the Theory of Statistical Estimation and Testing Hypotheses", Annals of Math. Stat., Vol. 10 (1939), pp. 299 - 326.

99. Wald, A.: Lectures on the Analysis of Variance and Covariance. Columbia University (1941).

100. Wald, A.: Notes on the Theory of Statistical Estimation and of Testing Hypotheses. Columbia University (1941).

101. Wald, A.: "Asymptotically Shortest Confidence Intervals", Annals of Math. Stat., Vol. 13, (1942), pp. 127 - 137.

102. Wald, A.: "Setting of Tolerance Limits when the Sample is Large", Annals of Math. Stat., Vol. 13, (1942), pp. 389 - 399.

103. Welsh, B. L.: "Some Problems in the Analysis of Regression among k samples of Two Variables". Biometrika, Vol. 27 (1935), pp. 145 - 160.

104. Whittaker, E. T. and Watson, G. N.: A Course in Modern Analysis, 4th ed., Cambridge University Press, 1927.

105. Whittaker, E. T. and Robinson, G.: The Calculus of Observations. London, Blackie and Son, 1932.

106. Widder, D. V., The Laplace Transform. Princeton University Press, 1941.

107. Wiener, N.: The Fourier Integral. Cambridge University Press, 1933.

108. Wilks, S. S.: "Certain Generalizations in the Analysis of Variance", Biometrika, Vol. 24 (1932), pp. 471 - 494.

109. Wilks, S. S.: "On the Sampling Distribution of the Multiple Correlation Coefficient" Annals of Math. Stat., Vol. 3 (1932), pp. 196 - 203.

110. Wilks, S. S.: "Moment-generating Operators for Determinants of Product Moments in Samples from a Normal System". Annals of Math., Vol. 35 (1934), pp. 312 - 340.

111. Wilks, S. S.: "On the Independence of k sets of Normally Distributed Statistical Variables", Econometrica, Vol. 3 (1935), pp. 309 - 326.

112. Wilks, S. S.: "The Likelihood Test of Independence in Contingency Tables", Annals of Math. Stat., Vol. 6 (1935), pp. 190 - 195.

113. Wilks, S. S.: "Shortest Average Confidence Intervals from Large Samples", Annals of Math. Stat., Vol. 9 (1938), pp. 166 - 175.

114. Wilks, S. S.: "Analysis of Variance and Covariance of Non-Orthogonal Data", Metron, Vol. XIII (1938), pp. 141 - 154.

115. Wilks, S. S.: "Determination of Sample Size for Setting Tolerance Limits", Annals of Math. Stat., Vol. 12 (1941), pp. 94 - 95.

116. Wilks, S. S.: "Statistical Prediction with Special Reference to the Problem of Tolerance Limits", Annals of Math. Stat., Vol. 13 (1942), pp. 400 - 409.

117. Wishart, J.: "The Generalized Product Moment Distribution in Samples from a Normal Multivariate Population", Biometrika, Vol. 20 A (1928), pp. 32 - 52.

118. Wishart, J. and Bartlett, M. S.: "The Generalized Product Moment Distribution in a Normal Distribution", Proc. Camb. Phil. Soc., Vol. 29 (1933), pp. 260 - 270.

119. Wishart, J. and Fisher, R. A.: "The Arrangement of Field Experiments and the Statistical Reduction of the Results". Imp. Bur., Soil. Sci., 1930. (Tech. Comm. 10.)

120. Wolfowitz, J.: "Additive Partition Functions and a Class of Statistical Hypotheses" Annals of Math. Stat., Vol. 13 (1942), pp. 247 - 279.

121. Yates, F.: "The Principles of Orthogonality and Confounding in Replicated Experiments", Jour. Agric. Science, Vol. 23 (1933), pp. 108 - 145.

122. Yates, F.: "Complex Experiments", Jour. Roy. Stat. Soc., Supplement, Vol. 2 (1935), pp. 181 - 247.

123. Yule, G. U.: An Introduction to the Theory of Statistics, 10th Ed., London, Griffin, 1936.

INDEX

Analysis of covariance, 195
  extension to several fixed variates, 199
Analysis of covariance table, 198
Analysis of variance, 176
  for incomplete lay-outs, 195
  for Graeco-Latin square, 192
  for Latin square, 189
  for randomized blocks, 180
  for two-way layout, 180
  for three-way layout, 186
  multivariate extension of, 250
Average outgoing quality limit, 223
Average quality protection, 223
Beta function, 75
Binomial distribution, 47
  Bernoulli case, 49
  moment generating function of, 48
  negative, 56
  Poisson case, 49
Binomial population, confidence limits of p in large samples from, 129
Borel-measurable point set, 10
Canonical correlation coefficient, 259
Canonical correlation coefficients, distribution of, in samples, 270
Canonical variate, 259
C. d. f. (cumulative distribution function), 5
Central limit theorem, 81
Characteristic equation of a variance-covariance matrix, 254
Characteristic function, 82
Characteristic roots
  of difference of two sample variance-covariance matrices, distribution of, 268
  of sample variance-covariance matrix, distribution of, 264
Chi square distribution, 102
  moment generating function of, 74
  moments of, 103
  reproductive property of, 105
Chi-square problem, Pearson's original, 217
Cochran's Theorem, 107
Complete additivity, law of, 6
Component quadratic forms, resolving quadratic into, 168
Conditional probability, 15
Conditional probability density function, 17
  for normal bivariate distribution, 62
  for normal multivariate distribution, 71
Confidence coefficient, 124
Confidence interval, 124
Confidence limits, 124
  from large samples, 127
  graphical representation of, 126
  of difference between means of two normal populations with same variance, 130
  of mean of normal population, 130
  of p in large samples from binomial population, 129
  of range of rectangular population, 123
  of ratio of variances of two normal population, 131
  of regression coefficients, 159
  of variance of normal population, 131
Confidence region, 132
Confounding, 186
Contagious distribution function, 55
Consistency of estimate, 133
Consumer's risk, 222
Contingency table, 214
  Chi-square test of independence in, 216
  likelihood ratio test for independence in, 220
Continuous distribution function, bivariate case, 10
  univariate case, 8
Convergence, stochastic, 81
Correlation coefficient, 32
  between two linear functions of random variables, 34
  canonical, 260
  canonical, distribution of, 270
  distribution of, 120

Correlation coefficient (con't)
multiple, 45
multiple, distribution of, in samples from normal multivariate population, 244
partial, 42
Covariance, 32
analysis of, 195
between two linear functions of random variables, 34
Critical region of a statistical test, 152
Cumulative distribution function,
bivariate case, 8
k-variate case, 11
continuous case, 10
continuous, univariate case, 8
discrete, bivariate case, 10
empirical, 2
mixed case, 11
postulates for, bivariate case, 9
postulates for, k-variate case, 12
postulates for, univariate case, 5
univariate case, 5
Curve fitting,
by maximum likelihood, 145
by moments, 145
Curvilinear regression, 166
Difference between two sample means, distribution of, 100
Difference of point sets, 5
Discrete distribution function,
bivariate case, 10
univariate case, 7
Disjoint point sets, 5
Distribution function,
binomial, 47
contagious, 55
cumulative, bivariate case, 8
cumulative, univariate case, 5
discrete, univariate case, 7
limiting, of maximum likelihood estimates in large samples, 138
marginal, 12
multinomial, 51
negative binomial, 54
normal bivariate, 59
normal multivariate, 63
Distribution function (con't)
normal or Gaussian, 56
of canonical correlation coefficients, 270
of characteristic roots of difference of sample variance-covariance matrices, 268
of characteristic roots of sample variance-covariance matrix, 264
of correlation coefficient, 120
of difference between means of two samples from a normal population, 100
of exponent in normal multivariate population, 104
of Fisher's z, 115
of Hotelling's generalized "Student" ratio, 238
of largest variate in sample, 91
of likelihood ratio for generalized "Student" statistical hypothesis, 238
of linear function of normally distributed variables, 99
of means in samples from a normal bivariate population, 100, 101
of means in samples from a normal multivariate population, 101
of median of sample, 91
of multiple correlation coefficient in samples from normal multivariate population, 244
of number of correct matchings in random matching, 210
of number of trials required to obtain a given number of "successes", 55
of order statistics, 90
of range of sample, 92
of regression coefficients, k fixed variates, 162
of regression coefficients, one fixed variate, 159
of runs, 201
of sample mean, limiting, in large samples, 81
of second order sample moments in samples from normal bivariate population, 116
of smallest variate in sample, 91
of Snedecor's F ratio, 114
of "Student's" ratio, 110
of sums of squares in samples from normal population, 102
of total number of runs, 203
Poisson, 52
Polya-Eggenberger, 56
Type I, 76

Distribution function (con't)
Type III, 72
Wishart, 120
Wishart, geometric derivation of, 227
Distribution functions, Pearson system of, 72
Efficiency of estimates, 134
Equality of means,
of normal populations, test of, 176
test for, in normal multivariate population, 238
Estimation,
by intervals, 122
by points, 133
Estimates,
consistency of, 133
efficiency of, 134
maximum likelihood, 136
optimum, 133
sufficiency of, 135
unbiased, 133
Expected value, 28
Factorial moments, 204
F distribution, Snedecor's, 114
Fiducial limits, 126
Finite population, sampling from, 83
Fisher's z distribution, 115
Fixed Variate, 16
Gamma function, 73
Gaussian distribution function, 56
Generalized sum of squares, 229
Graeco-Latin square, 190
analysis of variance for, 192
Gram-Charlier series, 76
Grouping, corrections for, 94
Harmonic analysis, 166
Hermite polynomials, 77
Hotelling's generalized "Student" ratio, 238
Incomplete lay-outs, 192
Independence,
linear, 160
in probability sense, 13
of mean and sum of squared deviations in samples from normal population, 108
of means and second order sample moments in samples from normal bivariate population, 120
Independence (con't)
of means and second order moments in samples from normal multivariate population, 120, 233
of sets of variates, test for, in normal multivariate population, 242
mutual, 14
statistical, 13
Inspection, sampling, 220
Interaction,
first order, 181
second order, 184
Jacobian of a transformation,
for k variables, 28
for two variables, 25
Joint moments of several random variables, 31
Lagrange multipliers, 97
Laplace transform, 38
Large numbers, law of, 50
Large samples, confidence limits from, 127
Largest variate in sample, distribution of, 91
Latin square, 186
analysis of variance for, 189
complete set of, 191
orthogonal, 190
Law of complete additivity, 6
Law of large numbers, 50
Least square regression function, 44
variance about, 44
Least squares, 43
Likelihood, 136
Likelihood ratio, 150
Likelihood ratio test, 150
in large samples, 151
for equality of means in normal multivariate populations, 238
for general linear regression statistical hypothesis for normal multivariate population, 247
for general normal regression statistical hypothesis, 170
for independence in contingency tables, 220
for independence of sets of variates in normal multivariate population, 242
for "Student" hypothesis, 150
for the statistical hypothesis that means in a normal multivariate population have specified values, 235

Limiting form of cumulative distribution function as determined by limiting form of moment generating function, 38

Linear combination of random variables, mean and variance of, 33

Linear combinations of random variables, covariance and correlation coefficient between, 34

Linear functions of normally distributed variables, distribution of, 99

Linear independence, 160

Linear regression, 40

generality of, 165

Linear regression statistical hypothesis, likelihood ratio test for, in normal multivariate populations, 247

Lot quality protection, 223

Marginal distribution function, 12

Matching theory,

for three or more decks of cards, 212

for two decks of cards, 208

Matrix, 63

Maximum likelihood, curve fitting by, 145

Maximum likelihood estimate, 136

Maximum likelihood estimates,

distribution of, in large samples, 138

of transformed parameters, 139

Mean of independent random variables, moment generating function of, 82

Mean value, 29

of linear function of random variables, 33

of sample mean, 80

of sample variance, 83

Means,

distribution of difference between, in samples from normal population, 100

distribution of, in samples from a normal bivariate population, 100

distribution of, in samples from a normal multivariate population, 101

Median of sample, distribution of, 91

M. g. f. (moment generating function), 36

Moment generating function, 36

of binomial distribution, 48

of Chi-square distribution, 74

of mean of independent random variables, 82

of multinomial distribution, 51

of negative binomial distribution, 54

Moment generating function (con't)

of normal bivariate distribution, 60

of normal distribution, 57

of normal multivariate distribution, 70

of Poisson distribution, 53

of second order moments in samples from a normal bivariate population, 118

Moment Problem, 35

Moments,

curve-fitting by, 145

factorial, 204

joint, of several random variables, 31

of a random variable, 30

Multinomial distribution, 51

moment generating function of, 51

Multiple correlation, 42

Multiple correlation coefficient, distribution of, in samples from normal multivariate population, 244

Negative binomial distribution, 54

moment generating function of, 54

Neyman-Pearson theory of statistical tests, 152

Normal bivariate distribution, 59

conditional probability density function for, 62

moment generating function of, 60

regression function for, 62

distribution of means in samples from, 101

distribution of second order moments in samples from, 116

independence of means and second order moments in samples from, 120

Normal distribution, 56

moment generating function of, 58

reproductive property of, 98

Normally distributed variables, distribution of linear function of, 99

Normal multivariate distribution, 63

conditional probability density function for, 71

distribution of exponent in, 104

distribution of subset of variables in, 68

moment generating function of, 70

regression function for, 71

variance-covariance matrix of, 68

Normal multivariate population,

distribution of means in samples from, 101

distribution of multiple correlation coefficient in samples from, 244

Normal multivariate population (con't)
distribution of second order moments in samples from a, 232
general linear regression statistical hypothesis for, 247
generalized "Student" test for, 234
independence of means and second order moments in samples from, 120, 233
test for independence of sets of variables in, 242
Normal multivariate populations, test for equality of means in, 238
Normal population,
distribution of means in samples from, 100
distribution of sums of squares in samples from, 102
independence of mean and sum of squared deviations in samples from, 108
Normal populations,
distribution of difference between means in samples from, 100
test of equality of means of several, 176
Normal regression, 157
fundamental theorem on testing hypothesis in, 170
k fixed variates, 160
one fixed variate, 157
Nuisance parameters, 130
Null hypothesis, 147
Optimum estimate, 133
Order statistics, 80
distribution theory of, 89
Ordering within samples, test for randomness of, 207
Parallelogram, area of, 7
Parallelotope, volume of, 228
Partial correlation, 45
coefficient, 45
P. d. f. (probability density function), 8
Pearson system of distribution functions, 72
Pearson's original Chi-square problem, 217
Point set,
difference, 5
product, 5
sum, 5
Poisson distribution, 58
moment generating function of, 63
Polya-Eggenberger distribution, 55
Population parameter, admissible set of values of, 147
Population parameters,
interval estimation of, 122
point estimation of, 133
Positive definite matrix, 63
Positive definite quadratic form,
k variables, 63
two variables, 59
Power curve of a statistical test, 154
Power of a statistical test, 152
Principal axes, 252
direction cosines of, 256
relative lengths of, 256
Principal components of a variance, 255
Probability, conditional, 15
Probability density function, 8
bivariate case, 11
conditional, 17
Probability element, 8
Probable error, 58
Producer's risk, 222
Product of point sets, 5
Quadratic form,
positive definite, 59
resolving, into component quadratic forms, 168
Quality control, statistical, 221
Quality limit, average outgoing, 223
Quality protection,
average, 223
lot, 223
Randomized blocks, 177
Randomness, 2
Randomness of ordering within samples, test for, 207
Random sample, definition of, 79
Random variable, definition of, 6
Range of sample, distribution of, 92
Rectangular population,
confidence limits of range of, 123
distribution of range in samples from a, 92
Regression, 40
Regression coefficient, 40